ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

Máster Universitario en Big Data: Tecnología y Analítica Avanzada

Uso de aprendizaje no supervisado para la predicción del *default*

Autora
Pilar Garzón Oliva

Dirigido por
Miriam Salcedo Montero

Madrid
Junio 2026

Pilar Garzón Oliva, declara bajo su responsabilidad, que el Proyecto con título **Uso de aprendizaje no supervisado para la predicción del *default*** presentado en la ETS de Ingeniería (ICAI) de la Universidad Pontificia Comillas en el curso académico 2025/26 es de su autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

Fdo.: ..... Fecha: 15 / 06 / 2026

Autoriza la entrega:

El Director del Proyecto

Miriam Salcedo Montero

Fdo.: ..... Fecha: 15 / 06 / 2026

V. B. del Coordinador de Proyectos

Carlos Morrás Ruiz-Falcó

Fdo.: ..... Fecha: 15 / 06 / 2026

A mi familia.

Resumen

El riesgo de crédito es uno de los riesgos centrales de la actividad bancaria. Su medición suele apoyarse en modelos de *credit scoring* entrenados con historiales de impago. Sin embargo, en productos nuevos, carteras recientes, exposiciones con pocos eventos de *default* u otras situaciones similares, el número de incumplimientos observados puede ser reducido. Este problema es especialmente relevante en los *Low Default Portfolios* (LDP), donde la escasez de impagos limita la capacidad de los modelos supervisados para aprender patrones fiables.

Este trabajo analiza si los métodos de aprendizaje no supervisado y de detección de novedades pueden aportar señal útil de riesgo cuando las observaciones de *default* son escasas. Para ello, se utiliza un conjunto de datos etiquetado de crédito al consumo correspondiente al período 2019–2020, compuesto por más de 1,5 millones de solicitudes y con una tasa de impago del 3,14 %.

La metodología combina modelos supervisados de referencia con técnicas no supervisadas y de detección de novedades. Los modelos supervisados permiten comprobar que el conjunto de datos contiene señal predictiva y sirven como base de comparación. Posteriormente, se evalúa si métodos como K-Prototypes, Gaussian Mixture Models o SVDD son capaces de identificar perfiles o segmentos con mayor riesgo sin depender directamente de las etiquetas de impago durante el ajuste del modelo.

Los resultados muestran que XGBoost obtiene la mayor capacidad discriminante cuando existe suficiente historial de *defaults*. No obstante, K-Prototypes identifica un segmento pequeño y estable con una tasa de impago superior a la media. Además, SVDD destaca entre los métodos de detección de novedades por mantener un comportamiento más estable cuando se reduce de forma extrema el número de impagos disponibles en el entrenamiento.

En conjunto, los resultados sugieren que los métodos no supervisados no sustituyen a los modelos supervisados cuando hay datos suficientes, pero sí pueden aportar valor como herramientas complementarias en carteras con pocos impagos.

Palabras clave: riesgo de crédito, LDP, aprendizaje no supervisado, *credit scoring*, K-Prototypes, SVDD.

Abstract

Credit risk is one of the core risks in banking activity. Its measurement usually relies on *credit scoring* models trained on historical default data. However, in new products, recent portfolios, exposures with few *default* events or similar situations, the number of observed defaults may be limited. This issue is especially relevant in *Low Default Portfolios* (LDP), where the scarcity of defaults limits the ability of supervised models to learn reliable patterns.

This thesis analyses whether unsupervised learning and novelty detection methods can provide useful risk signals when *default* observations are scarce. To do so, it uses a labelled consumer credit dataset from the 2019–2020 period, composed of more than 1.5 million applications and with a default rate of 3.14 %.

The methodology combines supervised benchmark models with unsupervised and novelty detection techniques. The supervised models are used to verify that the dataset contains predictive signal and to provide a reference point for comparison. Subsequently, the thesis evaluates whether methods such as K-Prototypes, Gaussian Mixture Models or SVDD are able to identify higher-risk profiles or segments without relying directly on default labels during model fitting.

The results show that XGBoost achieves the highest discriminative power when sufficient default history is available. However, K-Prototypes identifies a small and stable segment with a default rate above the portfolio average. In addition, SVDD stands out among the novelty detection methods because it maintains a more stable behaviour when the number of defaults available for training is severely reduced.

Overall, the results suggest that unsupervised methods do not replace supervised models when enough data are available, but they can provide value as complementary tools in portfolios with few observed defaults.

Keywords: credit risk, LDP, unsupervised learning, *credit scoring*, K-Prototypes, SVDD.

Índice general

Índice de figuras

Índice de tablas

Acrónimos

AUC	Area Under the Curve.
BSS	Brier Skill Score.
CRR	Capital Requirements Regulation.
DPD	Days Past Due.
EBA	European Banking Authority.
GMM	Gaussian Mixture Models.
IRB	Internal Ratings-Based Approach.
IV	Information Value.
KS	Kolmogorov-Smirnov statistic.
LDP	Low Default Portfolio.
PD	Probability of Default.
PSI	Population Stability Index.
SHAP	SHapley Additive exPlanations.
SVDD	Support Vector Data Description.
WOE	Weight of Evidence.

Símbolos

y_i	Variable objetivo de la solicitud i. Toma valor 1 si el cliente entra en default y 0 en caso contrario.
$\mathbf{x}_i$	Vector de variables explicativas de la solicitud i.
p	Tasa de default observada en la cartera.
K	Número de clusters o componentes utilizados por un modelo.
G	Coeficiente de Gini, utilizado para medir la capacidad discriminante del modelo.
ν	Parámetro de SVDD y One-Class SVM que controla la fracción esperada de observaciones fuera de la frontera aprendida.

Capítulo 1

Introducción

1.1. Motivación

La actividad bancaria se basa en asumir, medir y gestionar riesgos. Uno de los más relevantes es el riesgo de crédito, entendido como la posibilidad de que un cliente no devuelva total o parcialmente el dinero prestado. Esta situación puede darse, por ejemplo, en préstamos al consumo, tarjetas, hipotecas u otros productos de financiación. Para una entidad financiera, anticipar qué clientes tienen mayor probabilidad de impago es importante porque afecta directamente a la concesión de crédito, a las provisiones, al capital regulatorio y a la rentabilidad de la cartera.

Los modelos de *credit scoring* se utilizan precisamente para apoyar esa medición del riesgo. A partir de la información disponible en el momento de la solicitud, estos modelos asignan una puntuación de riesgo a cada cliente. Cuando el modelo está correctamente calibrado, esa puntuación puede transformarse en una probabilidad de impago, también denominada Probabilidad de Default (PD). Por tanto, conviene distinguir entre el *score*, que ordena a los clientes de mayor a menor riesgo, y la PD, que representa una estimación probabilística calibrada.

En condiciones normales, los modelos supervisados se ajustan con datos históricos en los que ya se conoce qué clientes terminaron pagando y cuáles entraron en *default*. Sin embargo, esta condición no siempre se cumple. En productos nuevos, carteras recientes, entidades con poca historia, exposiciones con pocos eventos de *default* u otras situaciones con información histórica limitada, el número de incumplimientos observados puede ser muy reducido. Cuando esto ocurre, los modelos supervisados pierden parte de la base estadística que necesitan para identificar patrones fiables. Este tipo de situaciones se conocen como *Low Default Portfolios* (LDP).

El problema tiene también una dimensión regulatoria. Los marcos de Basilea y la normativa europea requieren estimaciones robustas de la PD y períodos mínimos de

observación, especialmente cuando los modelos se utilizan en contextos de capital regulatorio o gestión prudencial del riesgo. En carteras LDP, cumplir estos requisitos puede resultar especialmente difícil porque apenas existen impagos sobre los que ajustar y validar los modelos. Como respuesta, los supervisores suelen exigir márgenes de conservadurismo adicionales. Estos márgenes refuerzan la prudencia, pero también pueden elevar las necesidades de capital sin mejorar necesariamente la capacidad predictiva del modelo.

En este contexto, resulta relevante explorar si la información disponible en la cartera puede aportar señal de riesgo incluso cuando las observaciones de *default* son escasas. En concreto, el comportamiento y el perfil observable de los clientes pueden contener patrones útiles. Por ejemplo, si un solicitante presenta características significativamente distintas a las de la mayoría de la cartera, es razonable analizar si su comportamiento financiero posterior también puede diferir. Esta idea motiva el uso de métodos de aprendizaje no supervisado y de detección de novedades como herramientas complementarias para identificar segmentos o perfiles de mayor riesgo sin depender directamente de la etiqueta de *default* durante el ajuste del modelo.

1.2. Pregunta de investigación e hipótesis

La pregunta central del trabajo es:

> *¿Pueden los métodos de aprendizaje no supervisado capturar señal útil de riesgo de crédito en una cartera con bajo número de impagos?*

Para responder a esta pregunta se plantean tres hipótesis:

H1 — Segmentación no supervisada: K-Prototypes es capaz de generar segmentos con perfiles de riesgo diferenciados sin utilizar la variable objetivo durante el ajuste del algoritmo.

H2 — Detección de anomalías y novedades: los clientes identificados como más atípicos por los algoritmos de detección de anomalías o novedades presentan una tasa de impago superior a la media de la cartera en el conjunto de validación.

H3 — Brecha frente al modelo supervisado: el mejor método sin uso directo del *default* como etiqueta de clasificación se sitúa por debajo del modelo supervisado en términos de Gini cuando existe suficiente historial de impagos, pero la brecha se reduce en escenarios LDP más extremos.

1.3. Contribuciones

Este trabajo aporta tres contribuciones principales. En primer lugar, desarrolla un *pipeline* reproducible que cubre la integración de 32 ficheros CSV, la depuración documentada de variables y el modelado final sobre más de 1,5 millones de solicitudes de crédito.

En segundo lugar, plantea una comparación entre modelos supervisados, métodos no supervisados y variantes de detección de novedades sobre una base de datos común y con métricas homogéneas. Esta comparación permite analizar qué parte de la señal de riesgo puede recuperarse sin utilizar directamente las etiquetas de impago como variable de clasificación.

En tercer lugar, se evalúa K-Prototypes como herramienta de segmentación de riesgo en datos mixtos [7]. Este método permite trabajar conjuntamente con variables numéricas y categóricas, lo que encaja bien con la naturaleza de las solicitudes de crédito. Además, se analiza la estabilidad temporal de la segmentación y se discute su posible utilidad como herramienta complementaria de seguimiento del riesgo de cartera.

1.4. Dataset

Los datos utilizados proceden de la competición *Home Credit – Credit Risk Model Stability* de Kaggle[1], publicada en 2024. El conjunto de datos contiene más de 1,5 millones de solicitudes de crédito al consumo correspondientes al período 2019–2020, distribuidas en 32 ficheros CSV. La información combina variables de la solicitud, historial interno de la entidad y registros procedentes de *bureaus* de crédito externos. La tasa de impago observada es del 3,14,

El *dataset* presenta dos características especialmente relevantes. La primera es su estructura relacional. La información está organizada en varios grupos temáticos con distintos niveles de granularidad, por lo que resulta necesario tomar decisiones de integración y agregación antes del modelado.

La segunda característica relevante es su dimensión temporal. Las solicitudes cubren el período 2019–2020, lo que permite analizar cómo cambian el volumen de solicitudes, la tasa de *default* y la estabilidad de algunas variables a lo largo del tiempo. En este trabajo, esta dimensión se utiliza principalmente para estudiar la evolución mensual del riesgo, calcular métricas de estabilidad como el *Population Stability Index* (PSI) y comprobar si algunas señales, como los *clusters* de K-Prototypes, mantienen un comportamiento consistente en distintos períodos.

[1] Ficha de la competición en Kaggle.

1.5. Estructura del trabajo

El trabajo se organiza en cinco capítulos. El Capítulo 2 revisa la literatura sobre *credit scoring*, LDP y métodos no supervisados aplicados al riesgo de crédito. El Capítulo 3 describe el proceso de integración, depuración y preparación de los datos. El Capítulo 4 presenta los modelos evaluados y compara sus resultados. El Capítulo 5 contrasta las hipótesis planteadas, resume los principales hallazgos y delimita los escenarios en los que los métodos no supervisados pueden aportar valor práctico. Finalmente, los anexos recogen el detalle del proceso de depuración.

Capítulo 2

Estado del arte

2.1. *Credit scoring* supervisado

El *credit scoring* tiene como objetivo estimar el riesgo de impago de un solicitante de crédito. En la práctica, esto se traduce en asignar a cada cliente una puntuación que permite ordenar las solicitudes de mayor a menor riesgo. Cuando el modelo está calibrado, esa puntuación puede transformarse en una Probabilidad de Default (PD). Desde un punto de vista estadístico, el problema suele plantearse como una clasificación binaria: a partir de un conjunto de características observadas en el momento de la solicitud, $\mathbf{x}_i$, se estima la probabilidad de que el cliente entre en *default*:

$$P(y_i = 1 \mid \mathbf{x}_i)$$

En este contexto, los modelos supervisados aprenden a partir de ejemplos históricos en los que se conoce si el cliente terminó pagando o entró en *default*. Por ello, su rendimiento depende en gran medida de la cantidad, calidad y representatividad de los *defaults* observados en el pasado.

2.1.1. Regresión Logística

La Regresión Logística es uno de los modelos de referencia en el sector financiero por su interpretabilidad [6]. Su objetivo es estimar la probabilidad de impago mediante la siguiente función:

$$P(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + e^{-(\beta_0 + \boldsymbol{\beta}^\top \mathbf{x}_i)}}$$

Una de sus principales ventajas es que sus coeficientes son fácilmente auditables. Cada

parámetro β_j mide el efecto de una variable sobre el *log-odds* de impago, manteniendo constantes el resto de variables. Esto facilita la explicación del modelo, su validación y su uso en entornos regulados.

Esta interpretabilidad tiene una contrapartida, la Regresión Logística asume relaciones lineales entre las variables explicativas y el riesgo de impago. En datos de crédito son frecuentes las interacciones entre variables y los efectos no lineales, por lo que esta hipótesis puede reducir su capacidad predictiva frente a modelos más flexibles.

Dentro de las *scorecards* bancarias, la transformación *Weight of Evidence* (WOE) y el *Information Value* (IV) son herramientas habituales para transformar y seleccionar variables. Un IV inferior a 0,02 suele interpretarse como ausencia de señal predictiva relevante, mientras que valores entre 0,02 y 0,10 indican una señal débil [14]. En contextos de bajo impago, este criterio debe interpretarse con cautela, ya que la escasez de *defaults* puede hacer que variables con información real presenten relaciones individuales aparentemente moderadas con la variable objetivo.

2.1.2. Árbol de Decisión

Los Árboles de Decisión son modelos supervisados que dividen progresivamente la muestra en grupos más homogéneos respecto a la variable objetivo. En cada nodo, el algoritmo selecciona la variable y el punto de corte que mejor separan a los clientes según su riesgo de impago [1].

Su principal ventaja es la interpretabilidad. La estructura del árbol permite seguir de forma sencilla las reglas de decisión que llevan a clasificar a un cliente como más o menos arriesgado. Por este motivo, los árboles resultan útiles como modelo de referencia y como herramienta exploratoria para entender relaciones no lineales simples entre variables.

Sin embargo, los Árboles de Decisión individuales también presentan limitaciones. Si no se controla su profundidad, pueden ajustarse demasiado a la muestra de entrenamiento y perder capacidad de generalización. Por ello, en este trabajo se utiliza un árbol con profundidad limitada, no como modelo final, sino como referencia interpretable frente a modelos más flexibles como XGBoost.

2.1.3. XGBoost

XGBoost [4] es un algoritmo basado en árboles de decisión que combina múltiples modelos débiles mediante *boosting* secuencial. En cada iteración, el algoritmo intenta corregir los errores cometidos por el conjunto de árboles anterior, optimizando de forma progresiva una función de pérdida.

Su principal ventaja frente a los modelos lineales es su capacidad para capturar relaciones no lineales e interacciones entre variables sin necesidad de especificarlas manualmente. Esta característica lo convierte en una herramienta especialmente potente en datos de crédito, donde el riesgo de impago puede depender de combinaciones complejas de variables financieras, demográficas y de comportamiento.

2.1.4. Interpretabilidad con SHAP

Una de las formas habituales de explicar las predicciones de un modelo basado en *boosting* es utilizar valores de Shapley. SHAP aplica esta lógica al aprendizaje automático y permite descomponer la salida del modelo en contribuciones asociadas a cada variable [9].

En este contexto, los valores SHAP ayudan a analizar qué variables empujan el *score* del modelo hacia mayor o menor riesgo. Su utilidad principal no es demostrar causalidad, sino aportar una explicación local y global del comportamiento del modelo, algo especialmente relevante cuando se emplean modelos complejos como XGBoost en un entorno bancario.

2.1.5. Evaluación de modelos de *credit scoring*

En problemas de riesgo de crédito, las clases suelen estar fuertemente desbalanceadas: la mayoría de clientes paga y solo una minoría entra en *default*. Por este motivo, la exactitud no es una métrica adecuada, ya que un modelo que predijera siempre que todos los clientes son solventes podría obtener un porcentaje de acierto elevado sin discriminar realmente el riesgo.

Por ello, en el sector financiero se utilizan métricas centradas en la capacidad de ordenación y discriminación del modelo:

- **Coeficiente de Gini**: se calcula como $G = 2 \times \mathrm{AUC} - 1$ y mide la capacidad del modelo para ordenar a los solicitantes de mayor a menor riesgo. En la práctica, un Gini en torno a 0,30 suele utilizarse como referencia mínima de capacidad discriminante útil [11], aunque este umbral no debe interpretarse como una regla universal y depende del producto, la cartera y el mercado.
- **Estadístico KS**: se define como la distancia máxima entre las distribuciones acumuladas de los *scores* de buenos y malos pagadores:

$$KS = \max_t |F_{\text{malos}}(t) - F_{\text{buenos}}(t)|$$

  Cuanto mayor sea esta separación, mayor será la capacidad del modelo para

distinguir entre ambos grupos.

- **Brier Score y Brier Skill Score** (BSS): el Brier Score mide el error cuadrático medio entre la probabilidad estimada y el resultado observado. Por ejemplo, si para una solicitud que finalmente entra en *default* el modelo asigna una probabilidad de 0,70, acumula un error de $(1 - 0,70)^2$. El BSS compara ese error con el de un modelo base que predice siempre la tasa media de *default*.

2.2. El problema de los *Low Default Portfolios*

En los *Low Default Portfolios*, la principal dificultad metodológica no está únicamente en que existan pocos impagos, sino en que ese número reducido de observaciones limita la estimación y validación de los modelos de riesgo.

El Comité de Supervisión Bancaria de Basilea reconoce esta dificultad dentro del enfoque IRB. Cuando existen pocos *defaults* históricos, las estimaciones de Probabilidad de Default (PD) presentan intervalos de confianza más amplios. Esto obliga a introducir márgenes de conservadurismo adicionales, lo que puede incrementar el capital regulatorio requerido sin mejorar necesariamente la capacidad predictiva del modelo.

En la práctica regulatoria europea [5], el problema suele aparecer cuando la PD observada es muy baja o cuando el número absoluto de *defaults* en el período de observación es reducido. En este trabajo, la tasa de impago del 3,14 % sitúa la cartera en una posición intermedia. Por un lado, el volumen total de observaciones permite construir modelos supervisados de referencia. Por otro, la clase de interés sigue siendo minoritaria, lo que permite analizar el comportamiento de los métodos no supervisados en un contexto de bajo *default* relativo.

El desequilibrio de clases afecta de forma distinta a los métodos supervisados y no supervisados. En los modelos supervisados, este problema puede abordarse mediante ponderación de clases, ajuste de umbrales o técnicas específicas de entrenamiento. En cambio, en los métodos no supervisados la dificultad es más estructural, ya que al no utilizar la variable objetivo, los algoritmos de *clustering* pueden distribuir los *defaults* entre varios grupos en lugar de concentrarlos en un segmento claramente más arriesgado. Por ello, resulta necesario comprobar empíricamente si estos métodos son capaces de capturar señal de riesgo sin acceder a las etiquetas de impago.

2.3. Métodos no supervisados en riesgo de crédito

Los métodos no supervisados no utilizan la etiqueta de *default* para aprender una regla de clasificación. En este trabajo se consideran dos familias: técnicas de *clustering*,

orientadas a construir segmentos de clientes, y métodos de detección de anomalías o novedades, orientados a identificar observaciones que se alejan del comportamiento habitual de la cartera.

2.3.1. *Clustering* para segmentación de riesgo

A diferencia de los modelos de clasificación, el *clustering* no predice directamente si un cliente entrará en impago. Su objetivo es identificar grupos de solicitantes con perfiles similares. Esta segmentación puede ser útil como paso previo al modelado supervisado, como herramienta de análisis exploratorio o como alternativa inicial en contextos donde no existen suficientes etiquetas de *default*.

K-Means es uno de los algoritmos de *clustering* más utilizados [10], pero tiene una limitación importante para datos de crédito: trabaja únicamente con variables numéricas. Las solicitudes de crédito suelen combinar información numérica y categórica, como ingresos, importes solicitados, tipo de contrato, estado civil o situación laboral. Transformar las variables categóricas en valores numéricos puede distorsionar las distancias y afectar negativamente a la calidad de los grupos.

K-Prototypes [7] aborda esta limitación extendiendo K-Means a datos mixtos. Para ello, combina la distancia euclídea en las variables numéricas con una medida de coincidencia en las variables categóricas:

$$E = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} \left[\underbrace{\sum_{m \in Q} (x_{im} - v_{km})^2}_{*\text{distancia numérica}} + \gamma \underbrace{\sum *m \in C \delta(x_{im}, v_{km})}_{\text{discrepancia categórica}} \right]$$

En esta expresión, Q representa el conjunto de variables numéricas, C el conjunto de variables categóricas, u_{ik} indica si la observación i pertenece al *cluster* k, v_{km} es el prototipo del *cluster* en la variable m, δ toma valor 1 cuando dos categorías no coinciden y γ pondera la parte categórica frente a la numérica. De esta forma, el algoritmo permite trabajar de manera más natural con la estructura real de los datos de crédito. En este trabajo, K-Prototypes se utiliza para comprobar si una segmentación construida sin etiquetas puede generar grupos con niveles de riesgo diferenciados.

2.3.2. Detección de anomalías y detección de novedades

Los métodos de detección de anomalías parten de la idea de que las observaciones inusuales pueden indicar comportamientos de mayor riesgo. En crédito, esta relación debe interpretarse con cautela. Un cliente con un perfil muy alejado del resto de la

cartera podría tener una probabilidad de impago distinta, pero no todos los clientes que impagan tienen por qué ser estadísticamente atípicos.

Por este motivo, la detección de anomalías no debe interpretarse como una solución directa al problema del impago, sino como una forma de explorar si existen perfiles atípicos asociados a mayor riesgo.

Isolation Forest [8] construye árboles de aislamiento aleatorios y mide cuántas particiones son necesarias para aislar cada observación. La lógica del método es que las observaciones normales suelen encontrarse en zonas más densas y requieren más particiones para quedar aisladas. En cambio, las observaciones más raras o extremas tienden a aislarse con menos particiones y reciben un mayor *score* de anomalía.

Local Outlier Factor [2] compara la densidad local de cada punto con la densidad de sus vecinos más próximos. A partir de esta comparación asigna a cada observación un factor de *outlier*. Si una observación se encuentra en una zona mucho menos densa que la de sus vecinos, obtiene un valor más alto y se considera más anómala dentro de su entorno local, aunque no necesariamente en el conjunto global de los datos.

One-Class SVM [13] adapta la lógica de las *Support Vector Machines* al caso en el que solo se dispone de observaciones consideradas normales. El método aprende una frontera que delimita la región habitual de los datos y considera más anómalas aquellas observaciones que quedan fuera de esa zona. En este trabajo se utiliza como detector de anomalías para comprobar si los clientes que se alejan del comportamiento normal de la cartera presentan mayor riesgo de impago.

El **SVDD** [15] sigue una lógica relacionada, pero plantea el problema como la búsqueda de una hiperesfera de mínimo volumen que contenga una fracción $(1 - \nu)$ de los datos normales. En su formulación clásica, esta frontera puede construirse mediante distintas funciones *kernel*. En este trabajo, la idea se utiliza para aprender el espacio de normalidad de los clientes solventes y detectar como más anómalas aquellas observaciones que se alejan de ese espacio aprendido.

Los **Gaussian Mixture Models** (GMM) modelan los datos como una combinación de K distribuciones gaussianas. Cada componente representa un perfil latente de clientes con características similares. Como detector de anomalías, el modelo considera más atípicas las observaciones con baja probabilidad bajo las distribuciones aprendidas.

Una variante especialmente relevante en contextos LDP es la **detección de novedades**. En este enfoque, el detector se ajusta exclusivamente con clientes solventes para aprender qué aspecto tiene un cliente que paga. Posteriormente, los clientes que se alejan de ese espacio aprendido se identifican como potencialmente anómalos. En este caso, el *target* no se utiliza como etiqueta de clasificación, sino únicamente para definir qué observaciones forman parte del conjunto de ajuste.

2.4. Estabilidad temporal

En riesgo de crédito, no basta con que un modelo funcione bien en una muestra concreta. También es necesario comprobar si mantiene su comportamiento cuando se aplica a datos posteriores. Esta estabilidad temporal es especialmente relevante en producción bancaria, donde las carteras, las condiciones económicas y el perfil de los solicitantes pueden cambiar con el tiempo.

El *Population Stability Index* (PSI) permite medir si la distribución de una variable cambia entre un período de referencia y un período posterior. Para ello, la variable se divide en intervalos y se compara el peso relativo de cada intervalo en ambos períodos:

$$\mathrm{PSI} = \sum_{i=1}^{n} (p_{\mathrm{actual},i} - p_{\mathrm{ref},i}) \cdot \ln\left(\frac{p_{\mathrm{actual},i}}{p_{\mathrm{ref},i}}\right)$$

donde $p_{\mathrm{ref},i}$ representa la proporción de observaciones en el intervalo i durante el período de referencia, y $p_{\mathrm{actual},i}$ representa esa misma proporción en el período de comparación.

De forma habitual, un PSI inferior a 0,10 se interpreta como una situación estable; valores entre 0,10 y 0,25 indican un cambio moderado; y valores superiores a 0,25 reflejan un cambio severo en la distribución. Estos umbrales deben entenderse como referencias prácticas, no como reglas absolutas, ya que su interpretación depende del tipo de variable, del tamaño de la muestra y del uso previsto del modelo.

Capítulo 3

Metodología

El trabajo se organiza en siete *notebooks* de Python que se ejecutan de forma secuencial. Los dos primeros se centran en la construcción y depuración del *dataset* de modelado. Los siguientes desarrollan los modelos supervisados, el análisis no supervisado, la detección de anomalías y novedades, y el experimento de sensibilidad LDP. Para garantizar la reproducibilidad, los experimentos utilizan una semilla aleatoria común cuando el algoritmo lo permite.

3.1. Arquitectura del dataset

Antes de aplicar cualquier técnica de modelado, es necesario transformar la información original en una matriz analítica única. Esta fase resulta especialmente importante porque el *dataset* no se presenta como una tabla plana, sino como un conjunto de ficheros relacionados entre sí y con distintos niveles de granularidad. Por ello, la arquitectura del dato condiciona las decisiones posteriores de integración, agregación y depuración.

3.1.1. Estructura relacional y clave de integración

El *dataset* de *Home Credit* está distribuido en 32 ficheros CSV con información procedente de distintos bloques: datos de la solicitud, historial interno de la entidad, información de créditos anteriores, registros fiscales, productos bancarios y *bureaux* de crédito externos. Esta estructura se aproxima a la de un entorno bancario real, donde la información relevante para evaluar el riesgo de un cliente suele estar repartida en diferentes tablas y debe consolidarse antes de tomar una decisión de riesgo.

La clave de integración es `case_id`, identificador único de cada solicitud. La tabla `train_base` se utiliza como tabla maestra, ya que contiene una fila por solicitud. A partir de ella se incorporan las demás tablas mediante *left joins*, de forma que no

se pierde ninguna solicitud aunque no exista información disponible en alguna tabla secundaria. En esos casos, los campos quedan como valores ausentes, lo que en algunos casos puede tener valor informativo propio.

La integración mantiene todas las solicitudes del conjunto base, sin pérdidas de observaciones. Tras transformar las variables de fecha en diferencias temporales respecto a `date_decision`, el *notebook* 01 genera el archivo `data_processed.parquet`, con 226 columnas.

3.1.2. Niveles de granularidad y protocolo de agregación

Uno de los principales retos del *dataset* es que no todas las tablas tienen la misma granularidad. Algunas contienen una única fila por solicitud, mientras que otras registran múltiples filas asociadas a un mismo `case_id`. Por ejemplo, puede existir una fila por cada crédito anterior del cliente o una fila por cada cuota de cada crédito.

Para integrarlas en una única tabla de modelado, las tablas con varias filas por solicitud se agregan hasta obtener una sola fila por `case_id`. La función de agregación se elige según el significado económico de cada variable:

- **Máximo**: para variables de comportamiento adverso, como los días de mora o retraso en pagos anteriores. El peor episodio histórico suele ser más informativo que el promedio.
- **Media**: para variables que reflejan el patrón habitual de comportamiento, como días medios de retraso o proporción de cuotas pagadas tarde.
- **Suma**: para conteos o importes acumulados, donde la magnitud total aporta más información que cada registro individual.
- **Primer valor disponible**: para variables estáticas del titular, como nivel educativo, estado familiar o tipo de vivienda.

De esta manera se convierte una estructura relacional compleja en una tabla analítica única, manteniendo el significado económico de las variables y evitando duplicar solicitudes durante el proceso de integración.

3.2. Análisis exploratorio

3.2.1. La variable objetivo y el desbalance de clases

La variable objetivo indica si una solicitud terminó en *default*. En el conjunto analizado, la tasa de *default* es del 3,14 %, lo que confirma un fuerte desbalance de clases. Esto

supone un ratio aproximado de 30,8 clientes solventes por cada cliente impagado.

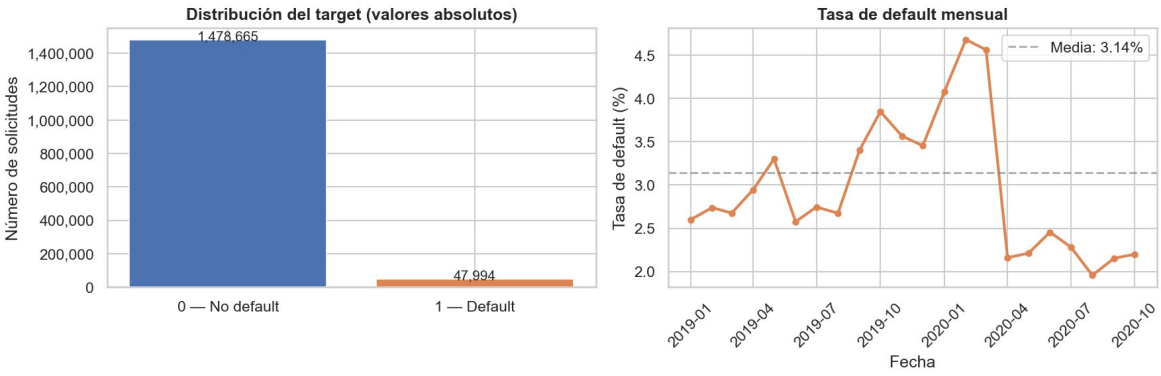


Figura 3.1: Distribución del *target* y evolución mensual de la tasa de *default*.

Este desbalance tiene varias implicaciones metodológicas. La primera afecta a las métricas de evaluación. La *accuracy* no resulta adecuada en este contexto, porque un modelo que clasificara a todos los clientes como solventes obtendría un porcentaje de acierto elevado sin haber aprendido a identificar el riesgo. Por ello, en los modelos posteriores se utilizan métricas orientadas a la discriminación y ordenación del riesgo, como el Gini, el KS y el Brier Skill Score.

La segunda implicación afecta a la interpretación de la señal estadística. Con una tasa de *default* reducida, incluso las variables realmente predictivas pueden mostrar correlaciones absolutas aparentemente moderadas. Por este motivo, las correlaciones con el *target* deben interpretarse teniendo en cuenta el fuerte desequilibrio de clases.

La tercera implicación es especialmente relevante para los métodos no supervisados. Si los clientes que impagan no presentan un perfil claramente diferenciado del resto de la cartera, las agrupaciones obtenidas pueden solaparse y los segmentos identificados pueden no concentrar claramente los impagos. Esta cuestión se analiza en el Capítulo 4.

3.2.2. Evolución temporal del default

Además de analizar la distribución global del *target*, se estudia la evolución mensual del volumen de solicitudes y de la tasa de *default*. Este análisis permite comprobar si el comportamiento de la cartera se mantiene estable o si existen cambios relevantes a lo largo del tiempo.

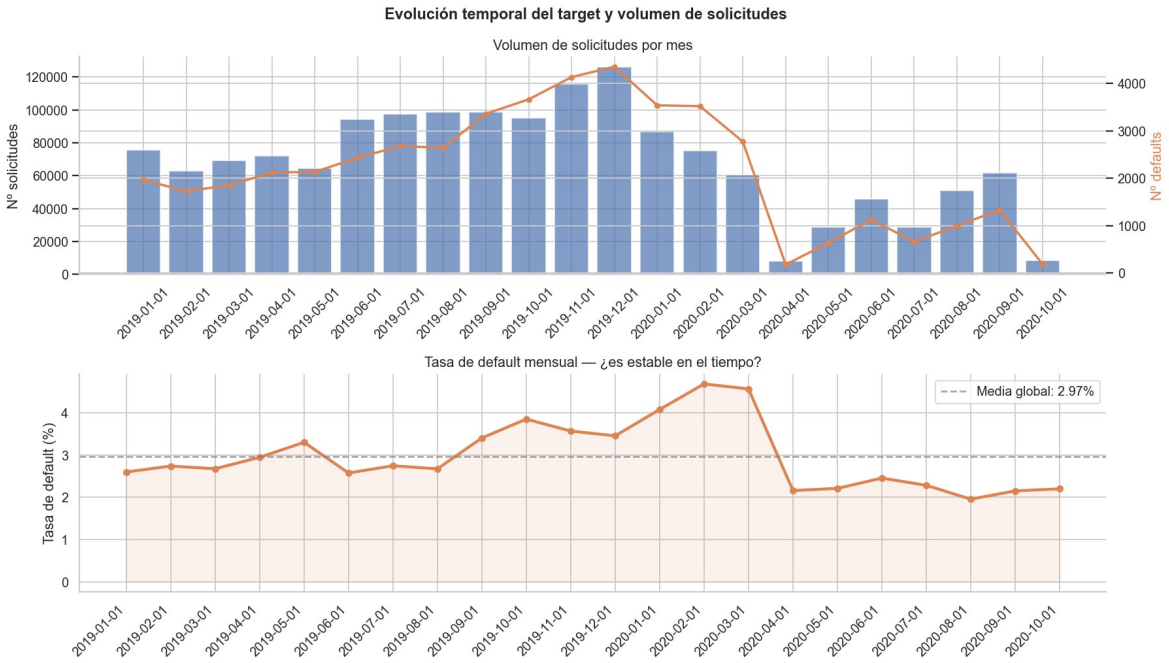


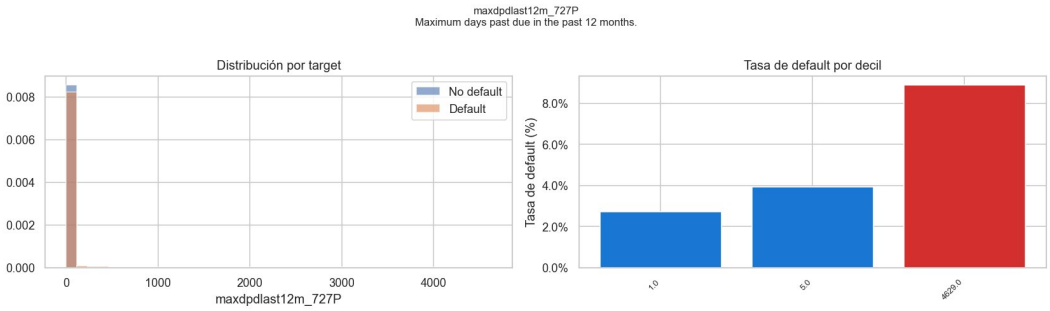
Figura 3.2: Evolución temporal del volumen de solicitudes y de la tasa de *default* mensual.

La evolución temporal es importante porque el período analizado incluye el año 2020, marcado por el inicio de la crisis del COVID-19. En carteras de crédito, este tipo de cambios puede afectar tanto al volumen de solicitudes como a la observación posterior de impagos. Por ello, el análisis temporal se utiliza como referencia para interpretar la estabilidad de los modelos y de las variables.

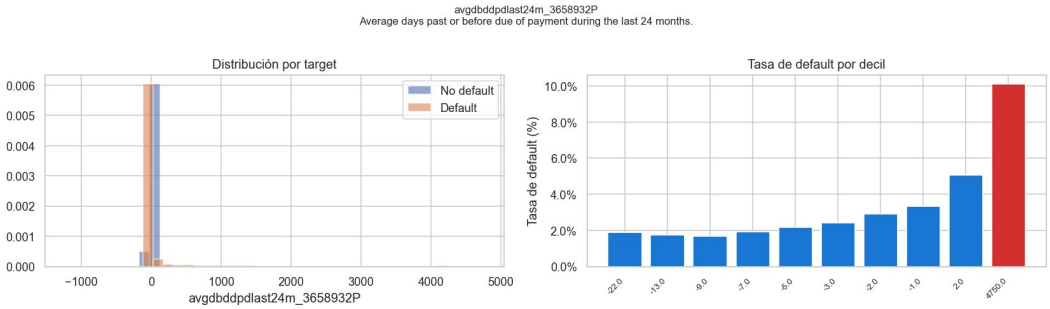
3.2.3. Análisis univariante de variables

Antes de construir los modelos, se analiza la relación individual de las variables más relevantes con el *target*. El objetivo de este análisis no es seleccionar todavía el modelo final, sino entender qué familias de variables contienen mayor señal de riesgo y si esa señal tiene sentido económico.

Las variables de mora histórica son las más relevantes. En particular, las variables relacionadas con días de mora, pagos realizados tarde y número de cuotas vencidas muestran una relación clara con el *default*. Este resultado es razonable: si un cliente ha tenido retrasos en pagos anteriores, puede tener una probabilidad mayor de volver a presentar problemas de pago.



(a) Mora máxima en los últimos 12 meses.



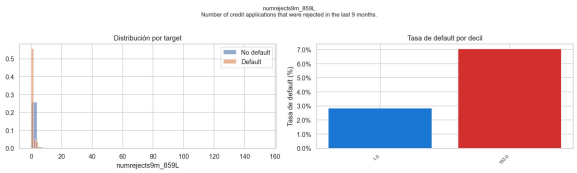
(b) Media de días de mora en los últimos 24 meses.

Figura 3.3: Variables de mora histórica.

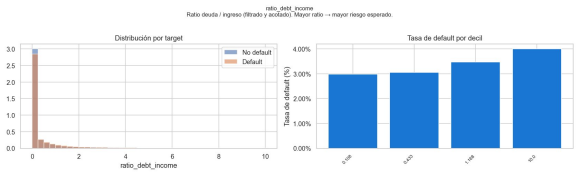
También resultan relevantes las variables de puntualidad de pago. Estas variables no solo identifican si el cliente ha tenido mora, sino que permiten diferenciar entre perfiles de pago más o menos disciplinados. Por ejemplo, un cliente que paga habitualmente con retraso presenta un comportamiento distinto al de un cliente que paga antes del vencimiento, aunque ambos puedan no haber registrado una mora grave.

Otra familia importante es la de historial de solicitudes y rechazos. Las variables que recogen rechazos recientes pueden interpretarse como una señal de presión financiera o búsqueda activa de liquidez. Si otras entidades han rechazado recientemente al cliente, esa información puede tener valor para evaluar su riesgo.

Finalmente, se incorporan variables relacionadas con la capacidad de pago. La deuda absoluta puede tener una interpretación limitada si no se compara con los ingresos del cliente. Por ello, en la fase de enriquecimiento se trabaja con el ratio deuda/ingreso, que aproxima la carga financiera relativa del solicitante.



(a) Rechazos en los últimos 9 meses.



(b) Ratio deuda/ingreso.

Figura 3.4: Variables de historial de solicitudes y capacidad de pago.

La matriz de correlaciones muestra que algunas variables de una misma familia están muy relacionadas entre sí. Esto ocurre especialmente en las variables de mora y puntualidad de pagos, donde distintos campos denotan aspectos próximos del comportamiento del cliente, aunque con ventanas temporales o umbrales diferentes. Esta redundancia se tiene en cuenta durante la depuración para evitar que una misma dimensión de riesgo pese varias veces en los métodos basados en distancias, como el *clustering*.

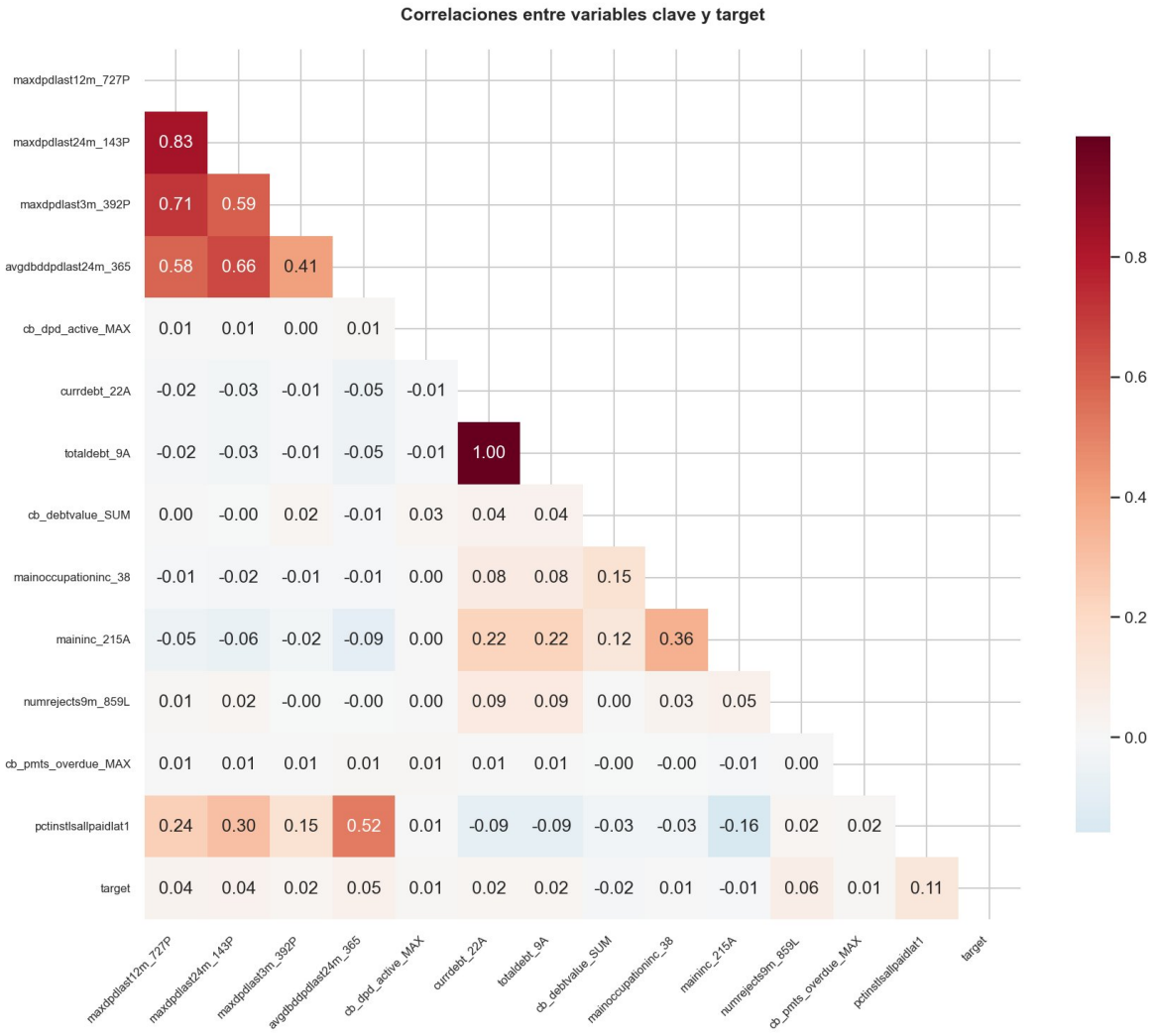


Figura 3.5: Correlaciones entre variables clave y el *target*.

3.2.4. Tratamiento de valores ausentes

El tratamiento de los valores ausentes es una parte central de la metodología. En este *dataset*, un valor ausente no siempre significa un error o falta de calidad del dato. En muchos casos, la ausencia de información refleja una característica del proceso de registro o de la cobertura disponible para ese cliente.

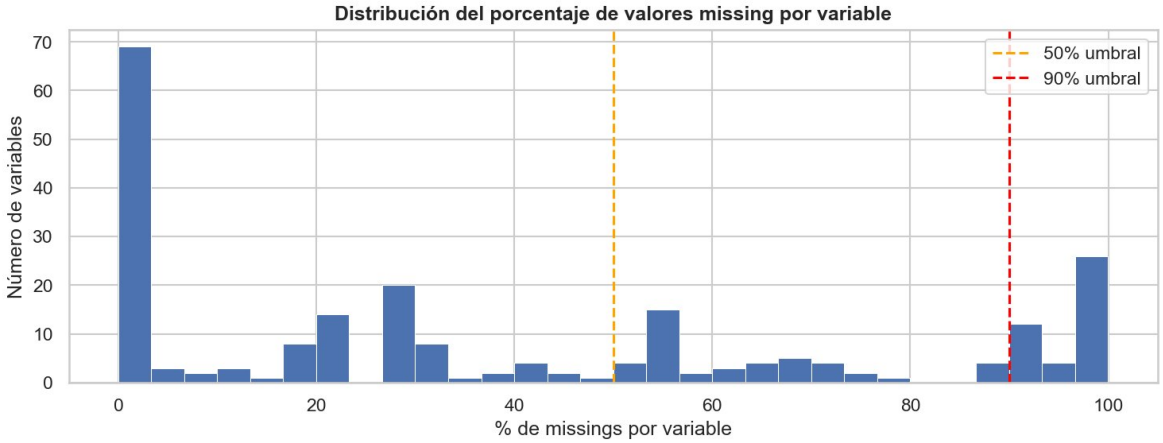


Figura 3.6: Distribución del porcentaje de valores ausentes por variable.

Se distinguen tres tipos de ausencia:

Ausencia informativa. En algunas variables, especialmente las procedentes de los ficheros de *bureau* de crédito, que el dato esté disponible o no puede ayudar a discriminar mejor el riesgo. En el análisis exploratorio se observa que los clientes con información disponible en determinadas variables externas presentan tasas de *default* distintas a las de los clientes sin dato. Por tanto, la ausencia de información no se trata simplemente como ruido, sino como una posible señal asociada a la disponibilidad de información sobre el cliente.

Ausencia operativa. En otros casos, el dato puede faltar por razones técnicas, de cobertura o de registro, sin que exista una interpretación económica clara. Estas variables se tratan posteriormente mediante imputación en los *pipelines* de modelado.

Ausencia por inaplicabilidad. Algunas variables corresponden a productos o situaciones que no aparecen en todos los casos del *dataset*. Cuando una variable no aporta variabilidad real, se elimina durante la depuración.

3.3. Depuración de variables

El *notebook* 02 parte del archivo `data_processed.parquet`, generado en el análisis exploratorio. Este archivo contiene más de 1,5 millones de solicitudes y 226 columnas. La depuración tiene como objetivo eliminar variables sin información útil, reducir redundancias y trabajar con nuevas variables que aporten interpretación económica.

El proceso se aplica de forma secuencial. Primero se eliminan variables claramente no informativas; después se revisan variables con cobertura extremadamente baja, variables con un porcentaje elevado de valores ausentes y familias de variables redundantes; finalmente, se crean indicadores de ausencia informativa y se revisan variables deriva-

das. El resultado final es el archivo `data_clean.parquet`, con más de 1,5 millones de solicitudes y 228 columnas.

3.3.1. Criterios de eliminación

Se aplican cinco criterios principales de depuración:

- **Variables constantes**: se eliminan cinco variables con varianza nula, es decir, variables que presentan el mismo valor en todos los registros válidos. Estas variables no pueden aportar capacidad predictiva porque no distinguen entre clientes.
- **Cobertura extremadamente baja**: se eliminan ocho variables con menos del 1 % de observaciones válidas. Este criterio se aplica a variables en las que casi todos los registros están ausentes y, por tanto, no existe muestra suficiente para estimar de forma estable su relación con el *target*. Es un criterio más restrictivo que el análisis general de valores ausentes, donde las variables con alta ausencia se revisan antes de decidir si se eliminan o se conservan.
- **Categóricas sin discriminancia**: se revisan las variables categóricas para comprobar si sus categorías presentan diferencias en la tasa de *default*. En la ejecución final no se descarta ninguna variable por este criterio, ya que no se detectan categóricas con *spread* de *default* igual a cero dentro del conjunto ya depurado.
- **Variables con muchos valores ausentes y sin señal**: se analizan las variables con más del 90 % de valores ausentes. No se eliminan automáticamente, porque una variable con alta ausencia puede seguir siendo informativa si la disponibilidad del dato discrimina riesgo. Por ello, solo se descartan aquellas que, además de tener muchos valores ausentes, no muestran señal suficiente en términos de correlación con el *target* o diferencia de tasa de *default* entre grupos. En la ejecución final se eliminan 22 variables por este criterio.
- **Variables equivalentes o redundantes**: se revisan variables que representan conceptos similares dentro del mismo *dataset*. Esto ocurre, por ejemplo, con variables procedentes de distintos ficheros, variables de la familia `clientscnt`, variables de mora DPD y porcentajes de cuotas pagadas tarde. En estos casos se conserva la versión con mayor cobertura, menor redundancia o interpretación más clara para el modelado.

3.3.2. Enriquecimiento del dataset

Tras la eliminación de variables, se incorporan elementos con valor económico e interpretativo.

En primer lugar, se crean indicadores de ausencia informativa. El criterio utilizado

consiste en identificar variables con un porcentaje relevante de valores ausentes y señal suficiente frente al *target*, medida mediante correlación o *spread* de *default*. En total, se crean 56 indicadores binarios de ausencia informativa. Estos indicadores permiten que los modelos capturen no solo el valor de una variable cuando existe, sino también el hecho de que dicha información no esté disponible.

En segundo lugar, se trabaja con tres variables derivadas de interpretación económica:

- `ratio_debt_income`: mide la relación entre la deuda actual y el ingreso principal del solicitante. Esta variable aproxima la capacidad de pago relativa del cliente.
- `tiene_mora_historica`: identifica si el cliente ha tenido algún episodio de mora histórica. Permite separar de forma sencilla a los clientes que nunca han presentado retrasos de aquellos que sí los han tenido.
- `mora_grave_historica`: identifica si el cliente ha tenido una mora superior a 30 días. Este umbral es especialmente relevante en riesgo de crédito, ya que señala un deterioro más severo del comportamiento de pago.

El análisis confirma que estas variables tienen sentido económico. Por ejemplo, los clientes con mora histórica presentan una tasa de *default* superior a los clientes sin mora, y la diferencia es todavía más clara cuando se considera la mora grave.

3.4. Dataset final de modelado y diseño experimental

El resultado de la depuración es el archivo `data_clean.parquet`, compuesto por 1.526.659 solicitudes y 228 columnas. La depuración no elimina observaciones, sino únicamente variables no informativas o redundantes. Por tanto, la tasa global de default se mantiene en el 3,14, %.

Este dataset limpio se utiliza como base común para todos los modelos posteriores. Mantener un único conjunto de datos de partida permite que la comparación entre modelos supervisados y no supervisados sea más homogénea. De esta forma, las diferencias observadas en los resultados se deben principalmente al enfoque de modelado y no a diferencias en la preparación previa de los datos.

El diseño experimental distingue tres bloques. En primer lugar, los modelos supervisados base permiten comprobar que el dataset contiene señal predictiva real. En segundo lugar, los métodos no supervisados y de detección de novedades evalúan qué parte de esa señal puede capturarse sin utilizar directamente el default como etiqueta de clasificación. En tercer lugar, el análisis de sensibilidad LDP reduce de forma progresiva el número de defaults disponibles en el entrenamiento para aproximar un escenario de bajo impago más extremo.

En las variantes de detección de novedades, el target se utiliza únicamente para definir

qué observaciones forman parte del conjunto de entrenamiento, normalmente clientes solventes. No se utiliza como etiqueta de clasificación dentro del algoritmo.

Capítulo 4

Resultados

Los resultados se presentan siguiendo la lógica del trabajo. Primero se analizan los modelos supervisados, que sirven como referencia y permiten comprobar que el *dataset* contiene señal predictiva. Después se estudia el *clustering* con K-Prototypes, que evalúa si es posible identificar segmentos de riesgo sin utilizar el *target* durante el ajuste del algoritmo. A continuación se presentan los detectores de anomalías y las variantes de detección de novedades. Por último, se analiza qué ocurre cuando se reduce el número de *defaults* disponibles en el entrenamiento, aproximando un escenario LDP más extremo.

4.1. Modelos supervisados de referencia

4.1.1. Regresión Logística y Árbol de Decisión

Los primeros modelos supervisados utilizados son la Regresión Logística y el Árbol de Decisión. Por un lado, estos modelos permiten comprobar que el *dataset* contiene señal predictiva real. Por otro, ofrecen una referencia interpretable frente a modelos más flexibles como XGBoost.

La Regresión Logística se ajusta con regularización L2 y ponderación de clases para compensar el fuerte desbalance entre clientes solventes e impagados. Este modelo resulta especialmente útil por su interpretabilidad, ya que sus coeficientes permiten identificar qué variables aumentan o reducen el riesgo estimado, manteniendo constantes el resto de predictores.

El Árbol de Decisión se configura con profundidad limitada para preservar cierta interpretabilidad y evitar un ajuste excesivo. Aunque este tipo de modelo puede capturar interacciones simples entre variables, su capacidad es más limitada que la de un *ensemble* de árboles.

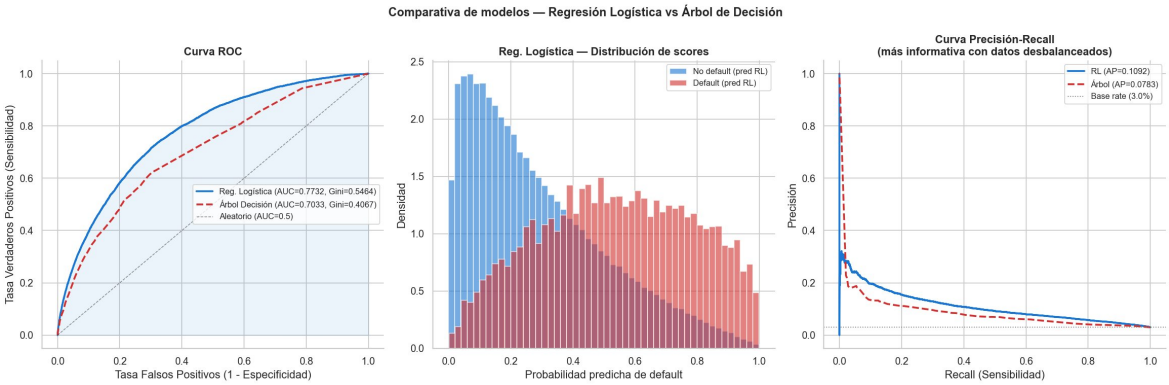


Figura 4.1: Comparativa entre Regresión Logística y Árbol de Decisión.

Los resultados muestran que ambos modelos capturan señal de riesgo, aunque con diferencias claras. La Regresión Logística alcanza un Gini de 0,546, mientras que el Árbol de Decisión obtiene un Gini de 0,407. La diferencia refleja que una estructura de árbol poco profunda no es suficiente para capturar toda la complejidad de las relaciones presentes en los datos.

La Figura 4.2 muestra los coeficientes estandarizados de la Regresión Logística. Al estar estandarizados, pueden compararse entre sí. Las variables asociadas a la carga financiera relativa, la mora histórica y el comportamiento de pago aparecen entre los factores de mayor peso. En sentido contrario, las variables que reflejan pagos anticipados o un historial de cumplimiento reducen el riesgo estimado.

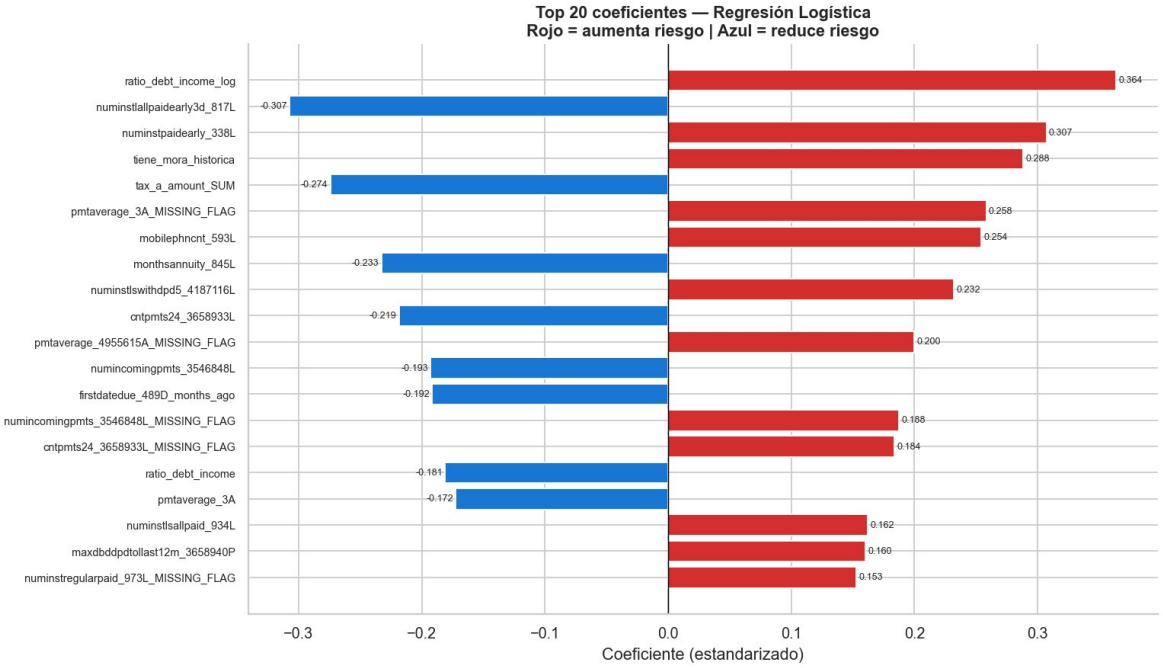


Figura 4.2: Top-20 coeficientes estandarizados de la Regresión Logística.

4.1.2. XGBoost

XGBoost se utiliza como modelo supervisado más flexible. A diferencia de la Regresión Logística, no exige que la relación entre las variables y el riesgo sea lineal. Al combinar múltiples árboles mediante *boosting*, puede capturar interacciones y efectos no lineales sin necesidad de especificarlos previamente.

El modelo se ajusta con ponderación de clases mediante `scale_pos_weight` para compensar el desbalance de la variable objetivo. El número de árboles se determina mediante *early stopping* sobre un conjunto de validación interno.

Los resultados del XGBoost confirman que el *dataset* contiene señal predictiva relevante. El modelo alcanza un AUC de 0,798, un Gini de 0,597 y un KS de 0,449, situándose como modelo supervisado de referencia del trabajo.

El objetivo principal de este modelo dentro del trabajo no es obtener una probabilidad de *default* perfectamente calibrada, sino ordenar a los clientes de mayor a menor riesgo. En este sentido, los valores de AUC, Gini y KS muestran que XGBoost tiene una buena capacidad de discriminación. No obstante, si sus salidas se quisieran interpretar directamente como probabilidades de impago para usos como provisiones, capital regulatorio o *pricing*, sería necesario aplicar una etapa específica de calibración.

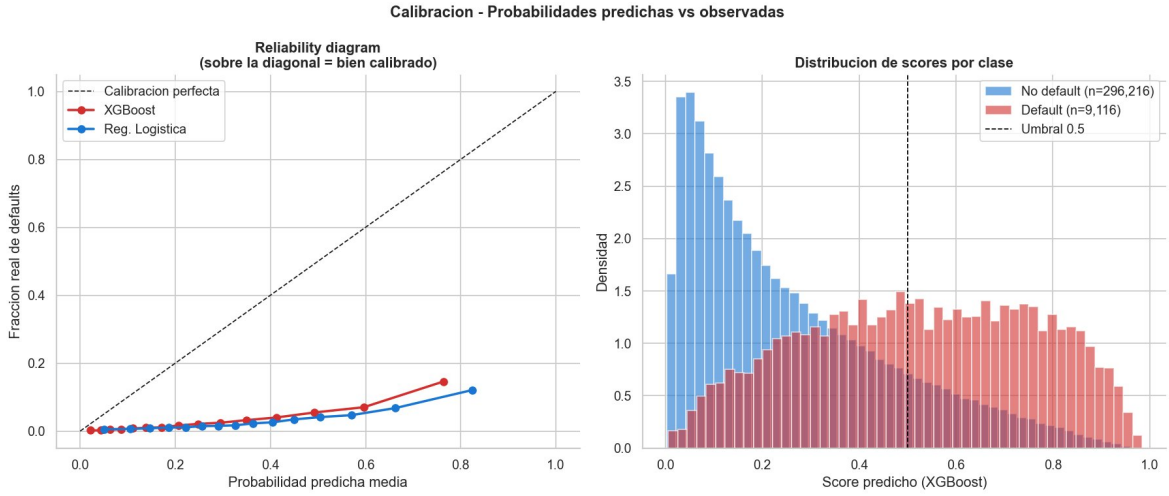


Figura 4.3: Calibración del XGBoost y la Regresión Logística.

4.1.3. Interpretabilidad con SHAP

Como se explicó en el Capítulo 2, una forma habitual de explicar las predicciones de modelos basados en *boosting* es utilizar valores SHAP. En este apartado se emplean como herramienta de interpretabilidad del modelo XGBoost. El objetivo no es establecer relaciones causales, sino analizar qué variables contribuyen en mayor medida al *score* del modelo y en qué dirección lo desplazan.

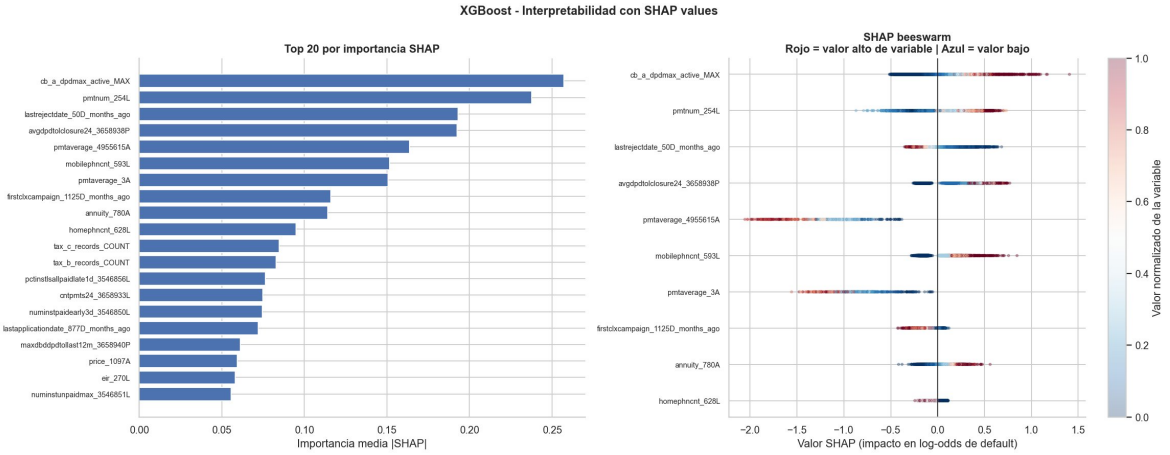


Figura 4.4: Interpretación global del XGBoost mediante valores SHAP.

Las variables más influyentes están relacionadas con la mora activa en el *bureau* externo, el historial de pagos, los rechazos recientes y la carga financiera. Este resultado es razonable, ya que el comportamiento pasado del cliente, tanto dentro como fuera de la entidad, influye de forma relevante en su riesgo futuro de impago.

El análisis individual de la Figura 4.5 muestra dos clientes *no-default* con perfiles de riesgo muy distintos. En un caso, las variables de pago actúan principalmente como señales protectoras. En el otro, la presencia de mora y rechazos recientes eleva el *score* de riesgo. Esta heterogeneidad dentro del grupo de clientes que no han impagado motiva el análisis no supervisado posterior: aunque dos clientes compartan el mismo resultado observado, su perfil de riesgo puede ser muy diferente.

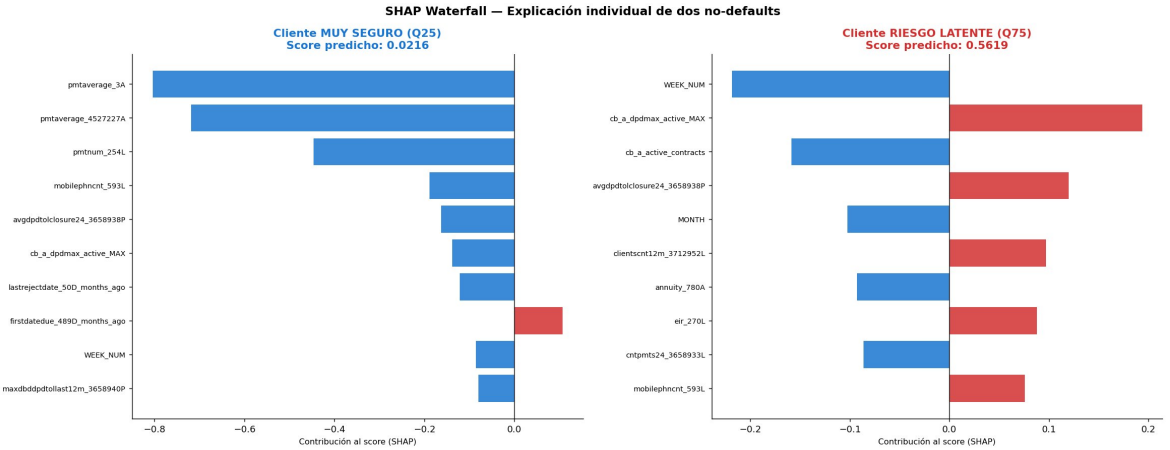


Figura 4.5: Contribuciones SHAP individuales para dos clientes *no-default* con perfiles de riesgo opuestos.

4.1.4. Estabilidad temporal

Además del rendimiento medio, se analiza la estabilidad temporal del modelo. Un modelo con buen Gini global puede no ser adecuado si su capacidad discriminante fluctúa demasiado entre meses o si las variables principales cambian de distribución de forma significativa.

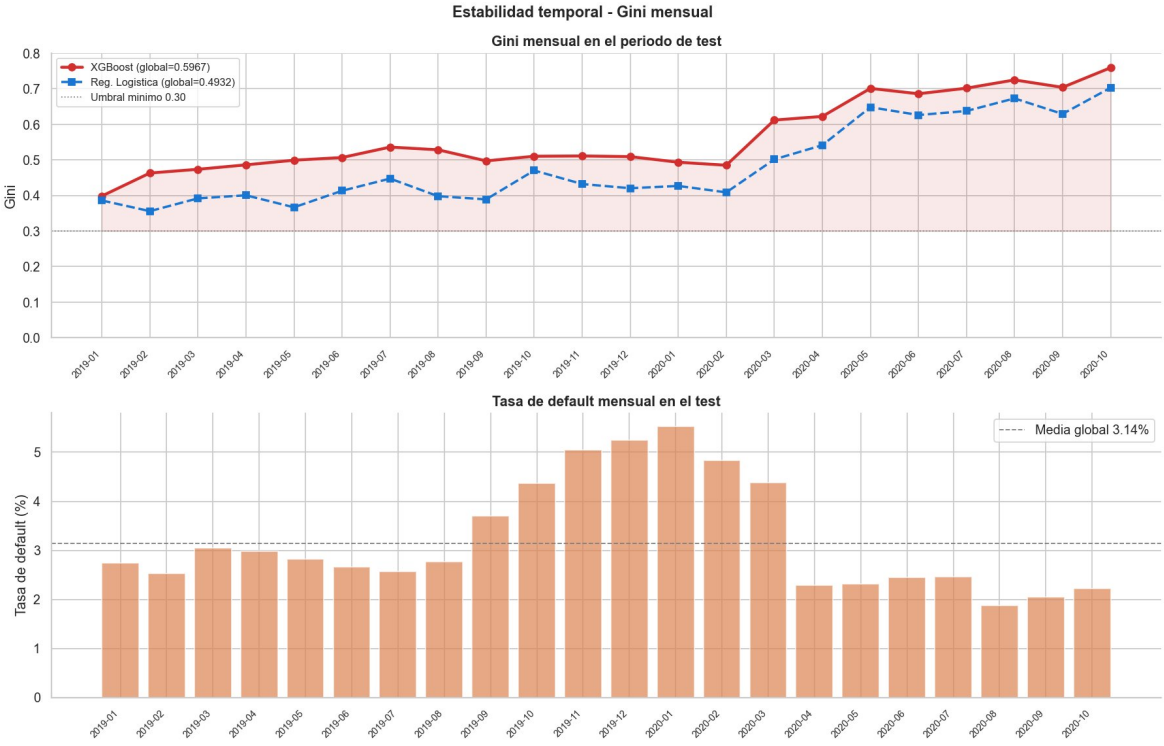


Figura 4.6: Gini mensual del XGBoost y la Regresión Logística.

El XGBoost mantiene un rendimiento claramente superior al azar, aunque presenta cierta variabilidad mensual. Esta variabilidad se interpreta en el contexto del período analizado, que incluye el inicio de la pandemia de COVID-19 y cambios relevantes en el volumen y composición de las solicitudes.

El PSI de las variables confirma que la mayoría presentan distribuciones estables. Las variables con PSI más elevado están asociadas principalmente a fechas o campañas comerciales, no a las dimensiones centrales de riesgo. En cambio, las variables de mora y comportamiento de pago, que concentran buena parte de la señal predictiva, se mantienen relativamente estables.

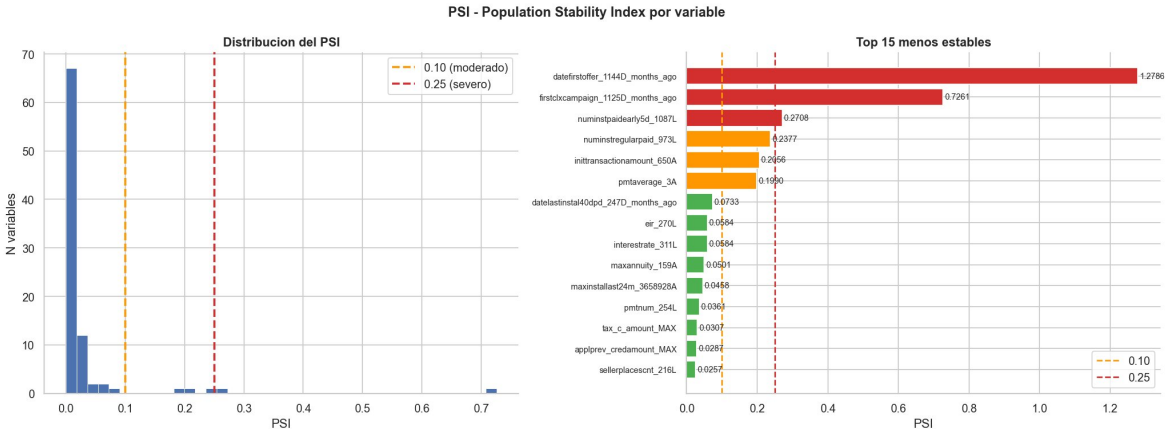


Figura 4.7: PSI de las variables numéricas utilizadas por el modelo.

Tabla 4.1: Resultados del *benchmark* supervisado.

Modelo	**AUC-ROC**	**Gini**	**KS**	**Brier**
Árbol de Decisión	0,703	0,407	0,318	—
Regresión Logística	0,773	0,546	0,414	0,169
XGBoost	0,798	**0,597**	0,449	0,115

El XGBoost se toma como referencia supervisada principal. A partir de este punto, la pregunta del trabajo es qué parte de esa señal puede recuperarse mediante métodos que no utilizan directamente el *default* como etiqueta de clasificación.

4.2. Clustering con K-Prototypes

El análisis SHAP muestra que los clientes *no-default* no forman un grupo homogéneo. Algunos presentan perfiles claramente solventes, mientras que otros tienen señales de riesgo latente. Esto lleva a evaluar si un método de *clustering* puede identificar grupos diferenciados sin utilizar la variable objetivo durante el ajuste del algoritmo.

K-Prototypes resulta adecuado para este problema porque permite trabajar con datos mixtos, combinando variables numéricas y categóricas. Esta característica es importante en solicitudes de crédito, donde conviven variables continuas, como ingresos o días de mora, con variables categóricas, como tipo de empleo, estado familiar o situación laboral.

4.2.1. Selección de K

Se evalúan distintos valores de K y se analizan métricas internas de *clustering*, como inercia, Silhouette [12] y Calinski-Harabasz [3]. Aunque aumentar el número de *clus-*

ters puede capturar más matices, también dificulta la interpretación y puede generar segmentos menos estables.

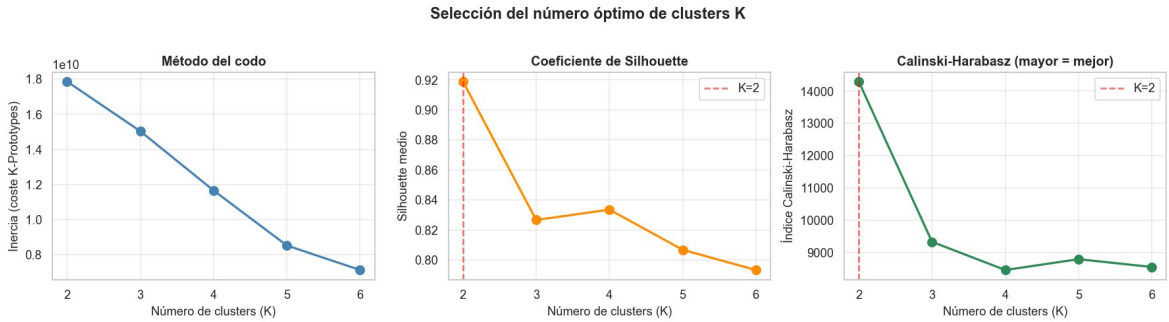


Figura 4.8: Selección del número de *clusters* en K-Prototypes.

La configuración principal se fija en $K = 2$. Esta elección permite obtener una segmentación sencilla, interpretable y con una separación clara entre un grupo mayoritario de bajo riesgo y un grupo minoritario de mayor riesgo.

4.2.2. Perfil de riesgo por cluster

El modelo final utiliza un conjunto reducido de variables con interpretación económica clara, como la mora histórica, la puntualidad de pago, la capacidad de pago relativa, el historial de solicitudes y las señales procedentes del *bureau*. El objetivo no es únicamente obtener una buena métrica interna de *clustering*, sino interpretar cada agrupación desde un punto de vista económico y comprobar si los segmentos presentan diferencias relevantes de riesgo dentro de la cartera.

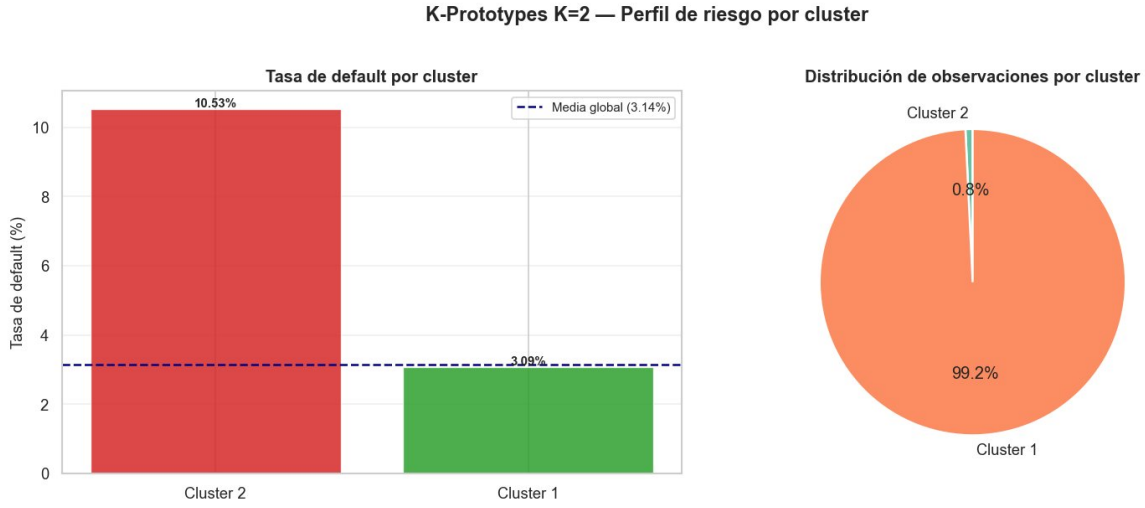


Figura 4.9: Perfil de riesgo por *cluster* en K-Prototypes.

Tabla 4.2: Perfil de riesgo por *cluster* — K-Prototypes $K = 2$.

Cluster	**N total**	**% cartera**	**DR *train***	**DR *test***	**Lift**
Solvente	1.492.101	97,7 %	2,96 %	2,39 %	0,9×
Alto riesgo	34.558	2,3 %	12,89 %	11,18 %	**4,1**×

El resultado más destacable es que el *cluster* de alto riesgo representa solo el 2,3 % de la cartera, pero alcanza una tasa de *default* del 11,18 % en el conjunto de *test*. Esto equivale a una tasa de impago 4,1 veces superior a la media. Lo más relevante es que esta separación se obtiene sin utilizar el *target* para construir la asignación final de las observaciones a *clusters*.

La variable que más separa ambos grupos está relacionada con la mora histórica. El *cluster* solvente concentra clientes con comportamiento de pago más estable, mientras que el *cluster* de alto riesgo agrupa perfiles con episodios de retraso más severos o persistentes.

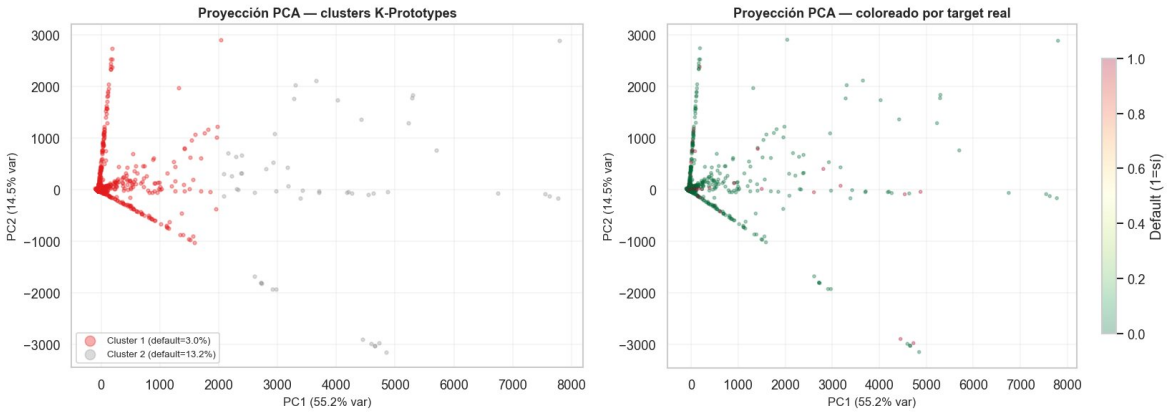


Figura 4.10: Proyección PCA de los *clusters* y del *target* real. La PCA se utiliza únicamente para representar en dos dimensiones una estructura de mayor dimensionalidad.

El Gini derivado del *clustering* como clasificador binario es 0,113, lo que sugiere una capacidad discriminante baja. No obstante, esta métrica no refleja completamente su valor. K-Prototypes no genera un *score* continuo de riesgo, sino una señal de segmentación. Su utilidad práctica está en identificar un grupo pequeño, estable y claramente más arriesgado que la media.

4.2.3. Estabilidad temporal del clustering

La estabilidad temporal es uno de los puntos fuertes del *clustering*. La tasa de *default* del *cluster* de alto riesgo se mantiene por encima de la media durante el período analizado, aunque con variaciones asociadas al menor tamaño del segmento.

Figura 4.11: Estabilidad temporal del *cluster* de alto riesgo.

El PSI máximo del indicador de *cluster* es 0,007, un valor muy inferior al umbral habitual de 0,10. Esto indica que la distribución de los *clusters* apenas cambia a lo largo del período analizado. Desde un punto de vista práctico, esta estabilidad hace que el *clustering* pueda ser útil como herramienta complementaria de seguimiento de cartera, aunque su capacidad discriminante sea menor que la de un modelo supervisado.

4.3. Detección de anomalías y novedades

La detección de anomalías se basa en la idea de que los clientes con perfiles poco habituales pueden presentar un riesgo de impago superior. Sin embargo, en crédito esta relación no siempre es directa. Un cliente puede acabar entrando en *default* aunque su perfil no sea especialmente atípico dentro de la cartera.

4.3.1. Detectores de anomalías estándar

Se evalúan detectores estándar, entendidos como modelos ajustados sobre toda la cartera sin utilizar etiquetas de *default* ni separar previamente clientes solventes e impagados. Estos modelos aprenden la estructura general de los datos y asignan mayor *score* de anomalía a las observaciones más alejadas de esa estructura.

Los resultados muestran que existe cierta señal, pero limitada. Los clientes con mayor *score* de anomalía tienden a presentar tasas de *default* algo superiores a la media, pero la separación entre *defaults* y *no-defaults* no es suficientemente fuerte para construir un modelo de *scoring* competitivo.

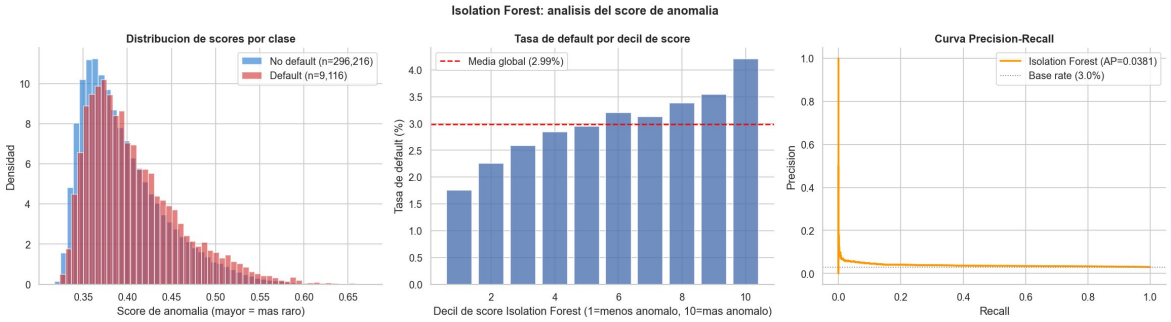


Figura 4.12: Diagnóstico del Isolation Forest estándar.

El solapamiento entre las distribuciones de *scores* de *defaults* y *no-defaults* indica que los impagos de esta cartera no se comportan como anomalías estadísticas puras. El riesgo existe, pero no aparece concentrado únicamente en observaciones extremas.

4.3.2. Detección de novedades

A continuación se evalúan variantes de detección de novedades. En este enfoque, el modelo se ajusta únicamente sobre clientes solventes. El objetivo es aprender con mayor precisión el espacio de normalidad y señalar como anómalos a los clientes que se alejan de él.

Este enfoque no es estrictamente no supervisado, porque utiliza el *target* para seleccionar qué observaciones forman parte del conjunto de ajuste. Sin embargo, no utiliza el *default* como etiqueta de clasificación directa. Por ello, resulta útil para medir cuánto valor aporta conocer, al menos, qué clientes han sido solventes.

Los resultados mejoran de forma consistente respecto a las variantes estándar. El Isolation Forest mejora de 0,206 a 0,221 en términos de Gini, mientras que GMM pasa de 0,220 a 0,234 cuando se ajusta únicamente sobre clientes solventes. El mejor resultado dentro de este bloque corresponde al SVDD, con un Gini de 0,251.

Aun así, la conclusión debe interpretarse con cautela. La detección de novedades permite recuperar cierta señal de riesgo, pero no separa con suficiente claridad a los clientes que entran en *default* de los que no lo hacen. Por tanto, su utilidad no está en sustituir a un modelo supervisado de *scoring*, sino en aportar una señal complementaria cuando las etiquetas de impago son escasas.

4.3.3. Gaussian Mixture Models y SVDD

Los *Gaussian Mixture Models* permiten representar la cartera como una mezcla de perfiles. Cada componente recoge un tipo de cliente con características similares. Esta granularidad puede ser útil para segmentación y *pricing*, ya que permite distinguir no

solo clientes de mayor riesgo, sino también grupos especialmente solventes.

El SVDD ofrece una interpretación geométrica distinta. Aprende una frontera alrededor del espacio ocupado por los clientes solventes y considera más anómalos a quienes quedan más alejados de ese espacio. En este trabajo, el SVDD obtiene el mayor Gini dentro del bloque de detección de novedades, por lo que se utiliza como principal referencia dentro de este grupo de métodos en el análisis de sensibilidad posterior.

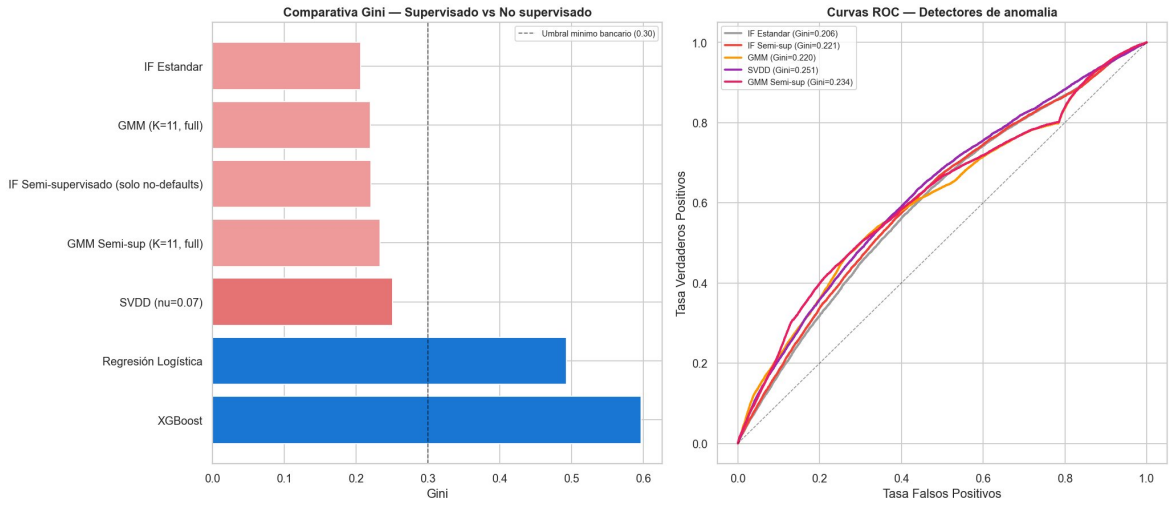


Figura 4.13: Comparativa de detectores avanzados.

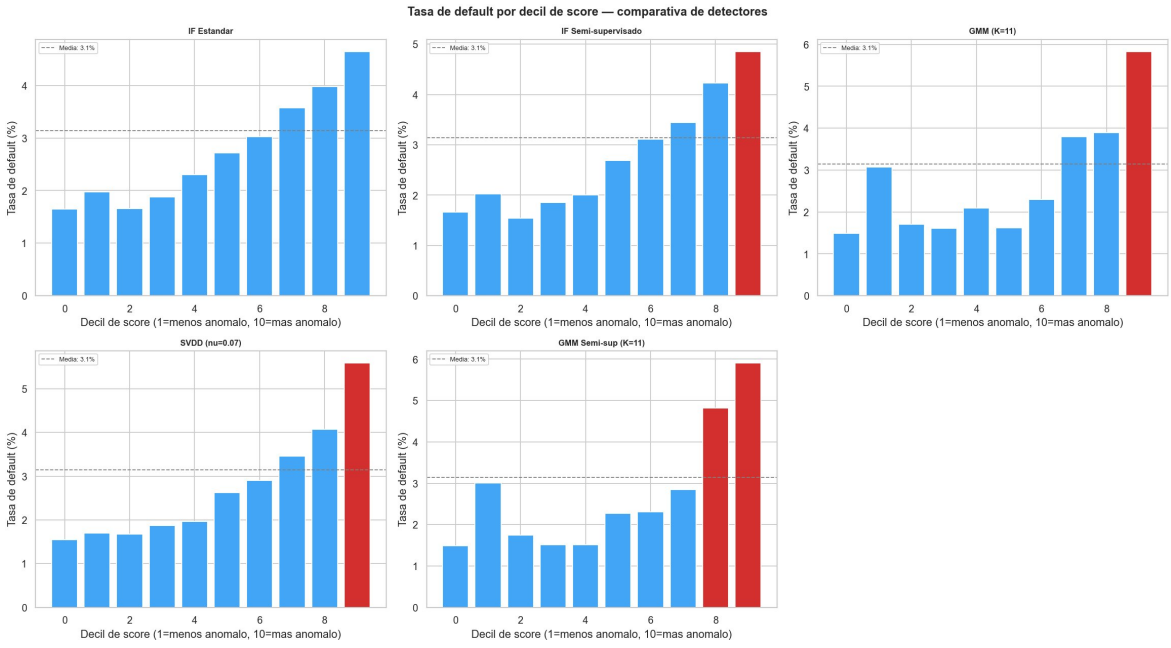


Figura 4.14: Tasa de *default* por decil de *score* para los detectores avanzados.

Aunque ninguno de los detectores alcanza el rendimiento del XGBoost, sí aportan señal útil. Su principal limitación es que la relación entre perfil atípico e impago no es directa en esta cartera.

4.4. Comparativa global

La Tabla 4.3 resume los resultados principales. Los modelos supervisados obtienen la mayor capacidad discriminante, con XGBoost como referencia. Entre los métodos sin uso directo del *default* como etiqueta de clasificación, el SVDD es el mejor detector individual, mientras que K-Prototypes destaca por su estabilidad y por la claridad de su segmentación.

Tabla 4.3: Comparativa global de modelos.

Modelo	**Tipo**	**Gini**
Árbol de Decisión	Supervisado	0,407
Regresión Logística	Supervisado	0,546
XGBoost	Supervisado	0,597
K-Prototypes $K = 2$	No supervisado	0,113
Isolation Forest	No supervisado	0,206
IF semi-supervisado	Detección de novedades	0,221
GMM	No supervisado	0,220
GMM semi-supervisado	Detección de novedades	0,234
SVDD	Detección de novedades	**0,251**

De esta comparación se extraen tres ideas. La primera es que las etiquetas de *default* tienen un valor claro, ya que cuando existen suficientes impagos los modelos supervisados obtienen mejores resultados. La segunda es que los métodos no supervisados pueden identificar segmentos o señales de riesgo sin entrenar un clasificador completo. La tercera es que la estabilidad debe interpretarse como un criterio complementario a la capacidad discriminante. K-Prototypes discrimina menos que SVDD en términos de Gini, pero produce una segmentación muy estable y fácil de interpretar. En concreto, el PSI máximo del indicador de *cluster* es 0,007, mientras que en XGBoost el PSI máximo de las variables analizadas es 0,018. Ambos valores están por debajo del umbral habitual de 0,10, aunque corresponden a análisis de estabilidad aplicados a objetos distintos.

4.5. Análisis de sensibilidad LDP

Los resultados anteriores se obtienen con un número suficiente de *defaults* para entrenar modelos supervisados de referencia. Sin embargo, en escenarios LDP más extremos puede haber tan pocos impagos observados que el modelo supervisado no disponga de ejemplos suficientes para aprender patrones estables.

Para analizar esta cuestión, se diseña un experimento de sensibilidad en el que se reduce progresivamente el número de *defaults* disponibles en el entrenamiento. En

concreto, se prueban siete escenarios, desde el conjunto de entrenamiento completo hasta un escenario extremo con 500 *defaults*. El conjunto de *test* se mantiene fijo y sin alterar, de modo que todos los modelos se evalúan sobre la misma cartera observada. Se comparan principalmente XGBoost, como modelo supervisado fuerte, y SVDD, como mejor alternativa basada en detección de novedades.

Tabla 4.4: Sensibilidad al número de *defaults* disponibles en entrenamiento.

Escenario	**XGBoost**	**SVDD**	$\Delta_{\text{XGB}-\text{SVDD}}$
Dataset completo	0,620	0,260	0,360
Escenario LDP extremo	0,291	0,261	0,029

La Tabla 4.4 resume los dos extremos del experimento, mientras que la Figura 4.15 muestra la evolución completa entre escenarios. Con suficientes *defaults*, XGBoost obtiene un rendimiento claramente superior. Sin embargo, al reducir los *defaults* hasta un escenario LDP extremo, su rendimiento se deteriora de forma notable. El SVDD, en cambio, permanece prácticamente estable. Como consecuencia, la diferencia entre ambos modelos, amplia en el conjunto completo, se reduce hasta ser casi inexistente.

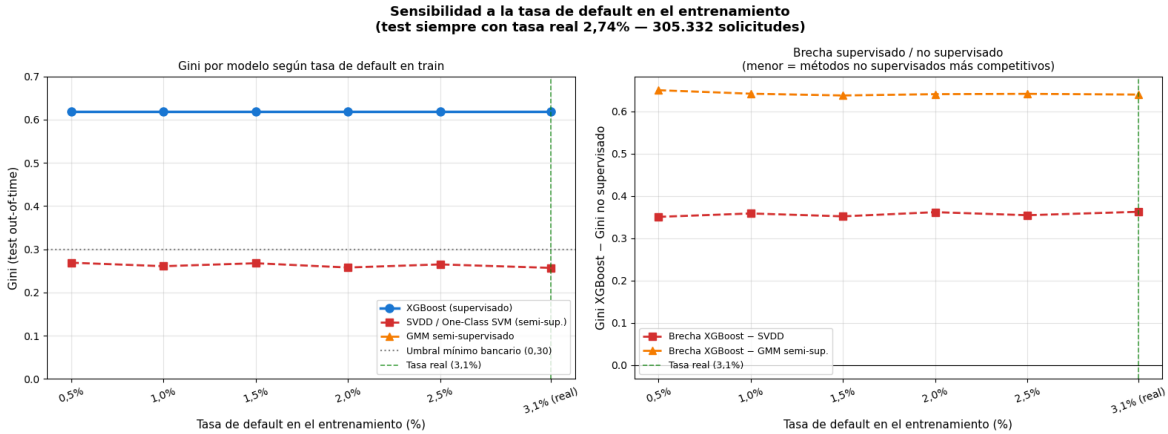


Figura 4.15: Sensibilidad al número de *defaults* disponibles en el entrenamiento.

Este resultado matiza la comparación global. Con suficiente historial de impagos, el modelo supervisado sigue siendo la alternativa más sólida. Sin embargo, cuando los *defaults* disponibles son muy escasos, su ventaja se reduce considerablemente y los métodos de detección de novedades pueden convertirse en una herramienta complementaria útil. Su valor no está en superar al aprendizaje supervisado en condiciones normales, sino en ofrecer una señal relativamente estable cuando las etiquetas de impago son insuficientes.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Contraste de hipótesis

El trabajo se organiza en torno a tres hipótesis operativas que permiten responder de forma estructurada a la pregunta de investigación.

La primera hipótesis (H1) analiza si K-Prototypes es capaz de generar segmentos con perfiles de riesgo diferenciados sin utilizar la variable objetivo durante el ajuste del algoritmo. Esta hipótesis se confirma. K-Prototypes con $K = 2$ identifica un segmento minoritario, equivalente al 2,3 % de la cartera, con una tasa de impago del 11,18 % en el conjunto de *test*. Esta tasa es 4,1 veces superior a la media y la distribución del indicador de *cluster* se mantiene estable a lo largo del período analizado, con un PSI máximo de 0,007. El resultado muestra que una segmentación no supervisada puede capturar información relevante de riesgo, aunque no sustituye a un modelo de *scoring* supervisado.

La segunda hipótesis (H2) evalúa si los detectores de anomalías y novedades identifican clientes con tasas de impago superiores a la media. Los resultados la apoyan parcialmente. Los detectores evaluados producen Gini positivos y los deciles más anómalos concentran más riesgo que la media de la cartera. Sin embargo, el efecto es moderado. El mejor resultado corresponde al SVDD, con un Gini de 0,251, por debajo del umbral de 0,30 utilizado como referencia orientativa de capacidad discriminante útil. Esto indica que los *defaults* no se comportan como anomalías estadísticas puras. La señal existe, pero no aparece concentrada únicamente en observaciones extremas.

La tercera hipótesis (H3) compara el mejor método sin uso directo del *default* como etiqueta de clasificación frente al modelo supervisado en términos de Gini. En el conjunto completo de datos, esta hipótesis se cumple claramente: XGBoost obtiene un Gini superior al de los métodos no supervisados y de detección de novedades. No obstante, el análisis de sensibilidad introduce un matiz importante. Cuando el número de

defaults disponibles en entrenamiento se reduce de forma extrema, el rendimiento del modelo supervisado disminuye de manera notable, mientras que el SVDD se mantiene prácticamente estable. En ese escenario, la diferencia entre ambos enfoques se reduce mucho, lo que refuerza la utilidad de la detección de novedades como herramienta complementaria en contextos LDP especialmente severos.

5.2. Hallazgos principales

El primer hallazgo relevante es que una segmentación no supervisada puede identificar grupos de riesgo diferenciados. K-Prototypes no alcanza la capacidad discriminante de un modelo supervisado, pero permite separar un grupo pequeño, interpretable y claramente más arriesgado que la media. Su principal valor no está en producir un *score* individual preciso, sino en ofrecer una estructura de segmentación estable y fácil de interpretar.

El segundo hallazgo es que los clientes que impagan en esta cartera no son necesariamente perfiles estadísticamente atípicos. Los detectores de anomalías parten de la idea de que las observaciones más inusuales pueden concentrar más riesgo. En este caso, esa idea solo se cumple parcialmente. Los clientes con mayor *score* de anomalía presentan más *defaults* que la media, pero el solapamiento entre *defaults* y *no-defaults* sigue siendo elevado. Esto sugiere que el riesgo de crédito está codificado en combinaciones complejas de variables, no únicamente en valores extremos.

El tercer hallazgo es que la detección de novedades mejora los resultados de los detectores estándar. Ajustar los modelos únicamente sobre clientes solventes permite aprender con mayor claridad el espacio de normalidad. Cuando un nuevo cliente se aleja de ese espacio, el modelo lo identifica como más anómalo. Esta mejora no convierte a estos métodos en modelos supervisados completos, pero muestra que la información sobre quién ha sido solvente tiene valor incluso cuando los *defaults* son escasos.

El cuarto hallazgo procede del análisis de sensibilidad LDP. La comparación entre XGBoost y SVDD cambia cuando se reduce el número absoluto de *defaults* disponibles. Con historial suficiente, XGBoost es claramente superior. Con muy pocos *defaults*, el modelo supervisado pierde gran parte de su ventaja, mientras que el SVDD mantiene un rendimiento estable. Esto refuerza una idea importante para los LDP: la dificultad no está solo en tener una tasa baja de *default*, sino en disponer de pocos impagos absolutos sobre los que aprender.

5.3. Implicaciones para la práctica bancaria

Cuando existe historial suficiente de impagos, los modelos supervisados siguen siendo la opción más adecuada. XGBoost ofrece la mayor capacidad discriminante, mientras que la Regresión Logística mantiene la ventaja de la interpretabilidad directa. En ambos casos, debe distinguirse entre la ordenación del riesgo y la estimación de probabilidades. Un modelo de *scoring* permite ordenar clientes de mayor a menor riesgo, pero sus salidas no pueden interpretarse directamente como probabilidades de *default*. Si se quieren utilizar para provisiones, *pricing* o capital regulatorio, es necesario incorporar una etapa específica de calibración y validación.

Cuando la cartera no tiene historial suficiente, K-Prototypes puede aportar una primera segmentación de riesgo sin necesidad de utilizar etiquetas de *default* durante el ajuste del algoritmo. Su principal ventaja no es producir un *score* individual preciso, sino identificar perfiles estructuralmente diferentes dentro de la cartera. En fases iniciales de un producto, mercado o cartera con poco historial, esta segmentación puede servir como herramienta de seguimiento y priorización.

Cuando se dispone de información fiable sobre clientes solventes, pero los *defaults* siguen siendo escasos, los métodos de detección de novedades son una alternativa complementaria razonable. SVDD y GMM semi-supervisado permiten aprender el espacio de normalidad y señalar clientes que se alejan de él. Estos métodos no sustituyen a un modelo supervisado cuando existe suficiente historial, pero pueden aportar señal en el arranque de carteras LDP.

Los *Gaussian Mixture Models* tienen, además, una utilidad potencial para la segmentación granular. Al representar la cartera como una mezcla de perfiles, pueden ayudar a identificar subgrupos de clientes con comportamientos diferenciados. Esta propiedad podría ser útil en políticas de seguimiento, análisis de cartera o estrategias de priorización.

5.4. Limitaciones

La primera limitación es computacional. Algunos detectores de anomalías se ajustan sobre submuestras por restricciones de memoria y tiempo de ejecución. Aunque las muestras utilizadas son suficientemente grandes para analizar patrones generales, el rendimiento sobre la totalidad de la cartera podría variar.

La segunda limitación tiene que ver con la explicabilidad. Aunque XGBoost ofrece buen rendimiento y SHAP permite interpretar sus predicciones, esta explicación no convierte al modelo en causal ni elimina la necesidad de validación experta. En un entorno bancario regulado, no basta con que un modelo discrimine bien: también debe

justificarse qué variables explican el riesgo, comprobar que la relación económica es coherente, revisar su estabilidad y validar su calibración antes de utilizarlo en decisiones de negocio, provisiones o capital regulatorio.

La tercera limitación afecta especialmente a los métodos no supervisados y de detección de novedades. Estos métodos producen segmentos o *scores* de anomalía, pero esas salidas son menos directamente traducibles a una probabilidad de *default*. Por tanto, su uso como modelos principales de decisión sería limitado sin una etapa adicional de validación, calibración y contraste experto. En este trabajo se interpretan principalmente como herramientas complementarias de segmentación, seguimiento o alerta temprana.

La cuarta limitación está relacionada con el *dataset*. Los datos proceden de mercados concretos y de una competición pública, por lo que no recogen necesariamente todas las particularidades de una entidad bancaria real. En mercados con mayor cobertura de *bureau*, políticas de concesión diferentes o perfiles de cliente distintos, la señal disponible para los métodos no supervisados podría cambiar.

La quinta limitación procede del período temporal analizado. El conjunto de datos incluye el inicio de la pandemia de COVID-19, que introduce cambios en el volumen de solicitudes, en la composición de la cartera y posiblemente en la observación de *defaults*. Aunque se analiza la estabilidad temporal, no se modela explícitamente el COVID-19 como un cambio de régimen.

Por último, todos los métodos evaluados tratan cada solicitud como una observación estática. Las variables de historial resumen el comportamiento previo mediante agregados, como máximos, medias o porcentajes, pero no capturan la secuencia completa de pagos. Un cliente que se recupera después de un episodio de mora y otro que se deteriora progresivamente pueden parecer similares si solo se observan estadísticos agregados. Esta es una de las limitaciones más importantes del trabajo.

5.5. Trabajo futuro

La primera línea de trabajo futuro consiste en incorporar explícitamente la dimensión temporal del comportamiento del cliente. En lugar de resumir el historial de pagos mediante máximos o medias, sería posible modelar la secuencia completa de pagos, retrasos y recuperaciones. Los modelos de supervivencia, las series temporales y las redes neuronales recurrentes podrían aportar una visión más rica del deterioro o mejora del perfil de riesgo.

Una segunda línea sería tratar el impacto del COVID-19 de forma específica. Separar el análisis en períodos pre-COVID y post-COVID, o incorporar variables macroeconómicas como covariables de calibración, permitiría distinguir mejor entre inestabilidad del modelo y cambios reales en el entorno económico.

Otra extensión relevante sería combinar las señales supervisadas y no supervisadas mediante un enfoque híbrido. Los *scores* de K-Prototypes, SVDD o GMM podrían añadirse como variables adicionales en un modelo supervisado. Si capturan información complementaria, podrían mejorar el rendimiento en segmentos concretos de la cartera o aumentar la estabilidad del modelo.

Finalmente, sería interesante profundizar en la calibración de los métodos no supervisados. Los *scores* de anomalía permiten ordenar clientes según su grado de rareza, pero no pueden interpretarse directamente como probabilidades de *default*. Por ello, una posible línea futura sería estudiar cómo transformar estos *scores* en estimaciones de PD calibradas, de forma que pudieran compararse mejor con los modelos supervisados y tener una interpretación más cercana al uso bancario real.

Apéndice A

Datos y preprocesamiento

A.1. Inventario de fuentes de datos

El *dataset* original de Home Credit se distribuye en 32 ficheros CSV dentro de la carpeta de entrenamiento. En el *Notebook* 01 se utiliza la tabla base de entrenamiento como punto de partida, ya que contiene una fila por solicitud y define el conjunto de observaciones del proyecto.

A partir de esta tabla se incorporan fuentes auxiliares con información de la solicitud, datos del cliente, historial interno, registros fiscales, productos bancarios y *bureau* de crédito externo. Cuando una fuente contiene varias filas para una misma solicitud, la información se agrega previamente hasta obtener una única fila por solicitud.

Los principales grupos de información utilizados son los siguientes:

- **Base de entrenamiento**: incluye el identificador de solicitud, la fecha de decisión y la variable objetivo.
- **Variables estáticas**: recogen información de la solicitud, ingresos, características del préstamo y perfil del cliente.
- ***Bureau* externo**: contiene información externa sobre créditos, saldos, pagos y episodios de mora.
- **Solicitudes anteriores**: recoge el historial de solicitudes previas del cliente con la entidad.
- **Información personal**: incluye datos personales del solicitante y de personas vinculadas a la solicitud.
- **Registros fiscales**: incorporan información procedente de declaraciones y registros asociados al cliente.
- **Productos bancarios**: incluyen información sobre depósitos, tarjetas, saldos y

otros productos financieros.

- **Cuotas históricas**: recogen el detalle de cuotas de créditos anteriores y el comportamiento de pago.

Las funciones de agregación se definen según el significado económico de cada variable. Se utilizan máximos para episodios de mora, sumas para importes acumulados, medias para comportamiento habitual y primeros valores disponibles para características estáticas.

A.2. Proceso de depuración

La integración realizada en el *Notebook* 01 genera un *dataset* procesado con 1.526.659 solicitudes y 226 columnas. A partir de ese archivo, el *Notebook* 02 aplica un proceso de depuración y enriquecimiento con el objetivo de eliminar variables poco informativas, reducir redundancias y conservar señales con interpretación económica.

El resultado final es el *dataset* limpio de modelado, compuesto por 1.526.659 solicitudes y 228 columnas. Aunque el número final de columnas es muy parecido al inicial, durante el proceso se eliminan variables sin utilidad y se crean nuevas variables que aportan información adicional.

El proceso de depuración aplicado en el *Notebook* 02 puede resumirse así:

- Se parte de un *dataset* integrado con 226 columnas.
- Se eliminan 5 variables constantes, ya que no presentan variabilidad útil para el modelo.
- Se eliminan 8 variables con menos del 1,
- No se eliminan variables categóricas por ausencia total de diferencia de riesgo entre categorías.
- Se eliminan 22 variables con más del 90,
- Se eliminan 12 variables redundantes de la familia `clientscnt`.
- Se eliminan 7 variables redundantes relacionadas con DPD.
- Se eliminan 2 variables intermedias de cuotas pagadas tarde.
- Se crean 56 indicadores binarios de ausencia informativa.
- Se incorporan o revisan variables derivadas relacionadas con capacidad de pago, mora histórica y mora grave.
- El *dataset* final queda compuesto por 228 columnas.

La diferencia entre las 226 columnas iniciales y las 228 finales se explica por el efecto

combinado de eliminaciones y nuevas variables. En particular, el *Notebook* 02 crea indicadores de ausencia informativa y trabaja con variables derivadas relacionadas con la capacidad de pago, la existencia de mora histórica y la presencia de mora grave.

A.3. Tratamiento de la ausencia informativa

En este *dataset*, un valor ausente no siempre debe interpretarse como un simple problema de calidad del dato. En algunas variables, especialmente las procedentes de fuentes externas como el *bureau*, que exista o no exista información disponible puede ayudar a discriminar mejor el riesgo.

Para comprobarlo, se revisa si la tasa de impago cambia entre los registros con dato disponible y los registros sin dato. Esta diferencia, o *spread* de *default*, permite detectar casos en los que la ausencia de información puede contener señal útil para el modelo.

Por este motivo, el *Notebook* 02 crea 56 indicadores binarios de ausencia informativa. Estos indicadores permiten que los modelos distingan entre clientes con información disponible y clientes sin información en determinadas fuentes. Su objetivo no es asumir que todo valor ausente implica más riesgo, sino evitar que una imputación simple elimine una diferencia potencialmente relevante entre grupos de clientes.

Estos indicadores no deben interpretarse como 56 variables individualmente fuertes. En muchos casos, su aportación individual es limitada. Su valor está en permitir que los modelos conserven información sobre la disponibilidad del dato, especialmente en fuentes externas y variables de historial.

A.4. Variables más discriminantes

El *Notebook* 03 calcula el *Information Value* de las variables numéricas para analizar su capacidad individual de separación entre clientes sin impago y clientes con impago. Las variables con mayor señal pertenecen principalmente a las familias de mora histórica, comportamiento de pago, *bureau* externo e historial reciente de solicitudes.

Las principales familias de variables con señal predictiva son:

- **Mora histórica**: recoge episodios previos de retraso en el pago y suele estar asociada a mayor riesgo futuro.
- **Comportamiento de pago**: resume la puntualidad del cliente en cuotas anteriores.
- ***Bureau* externo**: aporta información externa sobre exposición crediticia, pagos y deterioro previo.

- **Historial de solicitudes**: permite identificar señales recientes de presión crediticia o rechazos previos.
- **Capacidad de pago**: relaciona deuda, ingresos y carga financiera del solicitante.

El historial financiero del cliente y los episodios recientes de presión o deterioro crediticio concentran la mayor parte de la señal predictiva del *dataset*.

A.5. Decisiones sobre familias redundantes

Una parte importante de la depuración consiste en reducir familias de variables que miden conceptos muy similares. Esto es especialmente relevante para los métodos basados en distancias, como K-Prototypes, porque la presencia de muchas variables redundantes puede hacer que una misma dimensión de riesgo pese varias veces.

En el caso de las variables DPD, se eliminan variantes con ventanas temporales cortas, elevada correlación con otras variables ya conservadas o menor señal individual. También se revisan otras familias redundantes, como las variables `clientscnt` y los porcentajes de cuotas pagadas tarde.

Los criterios aplicados son los siguientes:

- En las variables DPD se conservan representantes con mayor señal, horizonte temporal más informativo o interpretación más clara.
- En la familia `clientscnt` se eliminan variables redundantes que miden conceptos muy próximos entre sí.
- En las variables de cuotas pagadas tarde se eliminan variantes intermedias cuando la información ya está recogida por otras variables más representativas.
- En variables con alta ausencia se conservan únicamente aquellas en las que la ausencia o disponibilidad del dato aporta señal frente a la variable objetivo.

Estas decisiones no buscan reducir variables mecánicamente, sino trabajar con un conjunto de información más limpio e interpretable de cara a los modelos posteriores.

Bibliografía

[1] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984.

[2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. *ACM SIGMOD Record*, 29(2):93–104, 2000.

[3] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[4] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[5] European Banking Authority. Guidelines on PD estimation, LGD estimation and the treatment of defaulted exposures. Technical Report EBA/GL/2017/16, EBA, 2017.

[6] David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. Wiley, Hoboken, NJ, 3 edition, 2013.

[7] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

[8] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 413–422, 2008.

[9] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[10] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[11] Nexialog Consulting. Scorecards backtesting. https://www.nexialog.com/wp-content/uploads/2023/10/Scorecards-Backtesting-Nexialog-Consulting-1.pdf, 2023. Accessed: 2026-06-10.

[12] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65, 1987.

[13] Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

[14] Naeem Siddiqi. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, Hoboken, NJ, 2006.

[15] David M. J. Tax and Robert P. W. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.