Master en Big Data

Master's final project

Integración de técnicas de *Machine Learning* en el Análisis de Supervivencia para la optimización de la LGD.

Escrito por
Julia González Parejo

Supervisado por
Gonzalo Ruiz Espinar

Madrid
June 2026

Resumen

Cuando un banco concede un préstamo asume el riesgo de que el prestatario no lo devuelva. Si eso ocurre, la entidad no solo pierde la posibilidad de cobrar las cuotas futuras, sino que debe estimar cuánto dinero va a recuperar finalmente a través de la liquidación de colateral, seguros u otros mecanismos. La fracción de la deuda que no se recupera es lo que se conoce como *Loss Given Default* o LGD, uno de los parámetros fundamentales que la regulación bancaria exige estimar con precisión para determinar cuánto capital debe mantener una entidad como colchón frente a pérdidas.

Estimar bien la LGD no es sencillo. El proceso de recuperación puede durar años, y en cualquier momento dado una parte de los préstamos en mora aún no ha llegado a su resolución final. Los modelos estadísticos tradicionales simplemente descartan esas observaciones incompletas o proyectan las recuperaciones futuras de estos préstamos para poder incluirlos, introduciendo incertidumbre adicional al modelo. Este trabajo propone aplicar el análisis de supervivencia, una técnica estadística habitualmente empleada en medicina para modelar el tiempo hasta la muerte de un paciente y especialmente diseñada para trabajar con este tipo de observaciones incompletas, al problema de la estimación de la LGD hipotecaria.

El estudio se desarrolla sobre una muestra del *Single-Family Loan-Level Dataset* publicado por Freddie Mac, que comprende hipotecas residenciales de Estados Unidos originadas entre 2005 y 2025. Tras un proceso detallado de depuración y construcción de la muestra, se trabaja con 28.502 episodios de *default*, definidos como ciclos continuos de morosidad que comienzan con la entrada en mora y terminan con una resolución definitiva o con el agotamiento del periodo de observación. La variable objetivo, la LGD, se construye a partir de los flujos de caja observados descontados al tipo de interés del propio préstamo, distinguiendo varios tipos de flujo según el estado del préstamo en cada mes del episodio.

Sobre esta muestra se comparan dos modelos de análisis de supervivencia: el modelo de riesgos proporcionales de Cox, un enfoque clásico y ampliamente utilizado, y el *Random Survival Forest*, una extensión basada en aprendizaje automático que permite capturar relaciones más complejas entre las variables. El objetivo no es sustituir ni optimizar los modelos de LGD tradicionales, sino estudiar si estas técnicas ofrecen una forma útil de tratar datos con procesos de recuperación largos y observaciones censuradas. Ambos modelos se evalúan mediante métricas de discriminación y calibración, y se complementan con técnicas de selección del horizonte temporal de predicción. Los resultados muestran que el enfoque de supervivencia permite incorporar los episodios censurados y ofrece predicciones competitivas frente a un modelo lineal de referencia, especialmente en el caso del *Random Survival Forest*. Además, se propone una extensión dinámica para actualizar la pérdida esperada de préstamos que ya llevan varios meses en mora, lo que abre la puerta a un uso más operativo del modelo en el seguimiento de carteras en default.

Abstract

When a bank grants a loan, it assumes the risk that the borrower may default. Should this occur, the institution not only loses the right to collect future payments, but must also estimate how much of the outstanding debt will ultimately be recovered through property sales, insurance claims, or other mechanisms. The fraction of the debt that is not recovered is known as the *Loss Given Default* (LGD), one of the fundamental parameters that banking regulation requires institutions to estimate accurately in order to determine the capital buffer they must hold against potential losses.

Estimating LGD accurately is not straightforward. The recovery process can take years, and at any given point in time a portion of the defaulted loans have not yet reached final resolution. Traditional statistical models either discard these incomplete observations or project the future recoveries of these loans in order to include them, introducing additional uncertainty into the model. This work proposes applying survival analysis, a statistical technique commonly used in medicine to model the time until a patient's death and specifically designed to handle this type of incomplete observation, to the problem of estimating mortgage LGD.

The study is conducted on a sample of the *Single-Family Loan-Level Dataset* published by Freddie Mac, comprising residential mortgages originated in the United States between 2005 and 2025. Following a detailed data cleaning and sample construction process, the analysis covers 28,502 default episodes , defined as continuous delinquency cycles that begin when the loan enters default and end either with a definitive resolution or with the exhaustion of the observation window. The target variable, the LGD, is constructed from the observed cash flows discounted at the loan's own interest rate, distinguishing several types of cash flow depending on the status of the loan in each month of the episode.

Two survival analysis models are compared on this sample: the Cox proportional hazards model, a classical and widely used approach, and the *Random Survival Forest*, a *machine learning* extension capable of capturing more complex relationships between variables. Both models are evaluated using discrimination and calibration metrics, and are complemented by techniques for selecting the prediction time horizon and *post-hoc* calibration. Additionally, an original methodological extension is proposed: the dynamic estimation of expected loss for loans that have already been in default for a given period of time, updating the prediction month by month as new information on the recovery process becomes available.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1.1 Contexto y motivación

La estabilidad del sistema financiero descansa, en gran medida, sobre la capacidad de las entidades de crédito para cuantificar con precisión las pérdidas a las que están expuestas. Entre los distintos tipos de riesgo que afrontan los bancos, el riesgo de crédito, definido como la posibilidad de que un prestatario incumpla sus obligaciones de pago, es históricamente el más relevante y el que mayor atención regulatoria ha recibido. Su correcta medición no es solo una exigencia técnica, sino sino la base sobre la que las entidades determinan las provisiones necesarias para cubrir las pérdidas que esperan sufrir en su cartera crediticia

El marco regulatorio internacional, conocido como Basilea II y, posteriormente, Basilea III, exige a las entidades cuantificar tanto la pérdida esperada de su cartera crediticia como el capital necesario para cubrir pérdidas inesperadas. En concreto, este trabajo se centra en el primer enfoque - la pérdida esperada - y dentro de él, en uno de sus tres componentes. La pérdida esperada (Expected Loss, EL) se expresa como:

$$EL = PD \times LGD \times EAD \tag{1.1}$$

La PD, o *Probability of Default*, recoge la probabilidad de que un prestatario incumpla sus obligaciones de pago en un horizonte temporal determinado, generalmente doce meses. La EAD, o *Exposure at Default*, mide el importe al que está expuesto el banco en el momento en que se produce el impago. La LGD, o *Loss Given Default*, representa la fracción de esa exposición que finalmente no se recupera una vez finalizado el proceso de recobro y es, por tanto, el parámetro que determina la severidad de la pérdida una vez que el impago se ha producido.

Las entidades que cuentan con la autorización del supervisor pueden estimar estos parámetros mediante modelos propios, lo que les permite calcular requerimientos de capital más ajustados al perfil real de riesgo de su cartera. Sin embargo, esta flexibilidad conlleva una responsabilidad: los modelos deben estar bien validados y reflejar fielmente el comportamiento esperado de las pérdidas, siendo objeto de supervisión continua por parte de los reguladores.

De los tres parámetros definidos, la PD ha sido históricamente el que más atención ha recibido. Mientras que esta ha sido objeto de estudio durante décadas, la LGD quedó durante mucho tiempo en un segundo plano. Esta comenzó a atraer atención a partir de los primeros documentos consultivos de Basilea II a principios de los 2000, cuando los bancos comenzaron a desarrollar modelos internos para su estimación. Ya entonces, con los primeros estudios de Schuermann (2004) y Altman, Resti y Sironi (2004), quedó de manifiesto que la LGD no era un parámetro sencillo de modelizar, no solo por la escasez o calidad de los datos disponibles, sino por la propia naturaleza estadística del fenómeno.

A diferencia de la PD, cuya naturaleza binaria la hace de por sí más sencilla de modelizar, la LGD presenta una distribución estadística que complica su modelización. En la práctica, los valores tienden a concentrarse en los extremos: hay préstamos que se recuperan casi en su totalidad y préstamos en los que la pérdida es prácticamente total, lo que da lugar a una distribución con frecuencia bimodal que los modelos de regresión lineal capturan con dificultad. A esto se suma que la LGD depende de factores muy heterogéneos entre sí, como el tipo de garantía, la vía de recobro utilizada, las condiciones macroeconómicas en el momento del default o las características propias del prestatario, lo que hace que su comportamiento sea difícil de generalizar.

Esta complejidad se ve además agravada por una dificultad de naturaleza estadística: el proceso de recuperación de un crédito en default puede extenderse durante un largo periodo de tiempo, lo que implica que, en cualquier momento dado, una parte significativa de las observaciones disponibles será incompleta. Los métodos clásicos de estimación tampoco están diseñados para manejar esta situación, lo cual motiva la búsqueda de otros enfoques para abordar el problema.

1.2 Objetivos del trabajo

En este trabajo se propone aplicar el análisis de supervivencia al problema de la estimación de la LGD, siguiendo una línea metodológica ya planteada en la literatura específica sobre modelización de LGD [1]. Esta técnica, muy utilizada en ámbitos como la medicina o la ingeniería, resulta especialmente adecuada cuando el evento que se quiere estudiar no se observa para todos los casos. En el contexto de la LGD, los préstamos cuyo proceso de recuperación aún no ha finalizado se incorporan al análisis,aprovechando la información parcial que contienen — cuánto se ha recuperado hasta el momento — sin necesidad de proyectar o simular su recuperación futura, como requieren los enfoques tradicionales.

A partir de esta idea, el trabajo tiene tres objetivos principales. El primero es construir una definición operativa de LGD a partir de los flujos de caja observados durante el proceso de recobro. El segundo es comparar un modelo tradicional de regresión con dos modelos de supervivencia: el modelo de Cox, como referencia clásica e interpretable, y el *Random Survival Forest* (RSF), que combina análisis de supervivencia con técnicas de aprendizaje automático. El tercero es analizar hasta qué punto estos modelos mejoran la predicción de la LGD, tanto en términos de error individual como de calibración agregada de cartera.

La comparación no pretende demostrar que el análisis de supervivencia sea siempre su-

perior a los enfoques tradicionales, sino evaluar si aporta una forma útil de aprovechar la dimensión temporal del problema y la información contenida en los episodios censurados. En particular, el uso del RSF permite explorar relaciones no lineales e interacciones entre variables, aunque a costa de una menor interpretabilidad frente a modelos más clásicos como el Cox PH o la regresión lineal.

Más allá de la estimación de la LGD, este trabajo plantea una extensión adicional: la predicción dinámica de la LGD residual para préstamos que ya llevan un determinado número de meses en mora. La idea es utilizar la curva de supervivencia estimada para actualizar la pérdida pendiente esperada a medida que avanza el *workout period*. Esta propuesta no busca cerrar un modelo definitivo de seguimiento mensual, sino mostrar que el marco de supervivencia también puede adaptarse a una visión más dinámica de la pérdida esperada, coherente con la necesidad de revisar periódicamente las estimaciones de riesgo a lo largo de la vida del préstamo.

1.3 Estructura del documento

El documento se organiza en seis capítulos. El Capítulo 1 introduce el contexto del problema, la motivación del trabajo y sus objetivos principales. El Capítulo 2 presenta el marco teórico, incluyendo los modelos tradicionales de regresión, los fundamentos del análisis de supervivencia, el modelo de Cox y los *Random Survival Forests*.

El Capítulo 3 describe la fuente de datos, la construcción de la muestra, la definición de los episodios de default y el cálculo de la LGD a partir de los flujos de caja observados. El Capítulo 4 recoge la metodología empírica: la construcción del dataset de supervivencia, la partición train/test, las métricas de evaluación, la selección de hiperparámetros y los métodos elección del horizonte temporal t^*.

El Capítulo 5 presenta los resultados obtenidos, comparando el modelo OLS, el Cox PH y el RSF, y analiza también la selección del horizonte temporal y la predicción dinámica de la LGD residual. Finalmente, el Capítulo 6 resume las principales conclusiones, expone las limitaciones del estudio y plantea posibles líneas de trabajo futuro.

Capítulo 2

Marco teórico

En este capítulo se presentan los fundamentos teóricos de las técnicas y métodos estadísticos utilizados a lo largo del trabajo, desde enfoques más clásicos como el de regresión que servirán como referencia, hasta el análisis de supervivencia y su integración con *Machine Learning* a través de los *Random Survival Forests*.

Cabe señalar que en la práctica bancaria la estimación de la LGD se apoya mayoritariamente en modelos de regresión, siendo la aplicación de técnicas más avanzadas todavía poco extendida. Este capítulo sienta las bases teóricas necesarias para entender tanto la construcción de los modelos como la interpretación de sus resultados en el contexto del problema que se presenta [2].

2.1 Enfoque tradicional: modelos basados en regresión

Los modelos de regresión constituyen el enfoque más utilizado en la práctica bancaria para la estimación de la LGD especialmente por su sencillez e interpretabilidad, cualidades muy valoradas en un entorno regulatorio donde los modelos deben poder explicarse y justificarse ante los supervisores.

El modelo más directo es la regresión lineal, que estima la LGD de cada préstamo como una combinación lineal de sus características:

$$LGD(a) = x(a)'\beta + \varepsilon \tag{2.1}$$

Donde $x(a)$ es el vector de variables explicativas, β el vector de coeficientes a estimar y ε el término de error. Los coeficientes se obtienen minimizando la suma de errores cuadráticos sobre el conjunto de entrenamiento.

Una alternativa dentro de los modelos de regresión es la regresión logística, que en lugar de predecir directamente la LGD estima la probabilidad de que la pérdida sea alta o baja en función de un umbral predefinido. La probabilidad de que un préstamo a pertenezca a uno de los dos grupos, por ejemplo al grupo donde la pérdida es baja, se estima mediante:

$$\pi(a) = \frac{\exp\left(x(a)'\beta\right)}{1 + \exp\left(x(a)'\beta\right)} \tag{2.2}$$

Al igual que en la ecuación 2.1, $x(a)$ recoge las variables explicativas del préstamo y β los coeficientes estimados por el modelo. Para construir los dos grupos, utilizamos el umbral θ previamente definido, que permite clasificar los préstamos históricos en operaciones de pérdida baja o pérdida alta.

La predicción final se obtiene combinando los valores medios observados de la LGD en cada grupo, ponderados por esa probabilidad:

$$\hat{L}(a) = \pi(a) \cdot \mu_{\text{bajo}} + (1 - \pi(a)) \cdot \mu_{\text{alto}} \tag{2.3}$$

Donde μ_{bajo} denota el valor medio de la LGD observado entre los préstamos clasificados por debajo del umbral θ, y μ_{alto} el valor medio de la LGD observado entre aquellos clasificados por encima de dicho umbral. De este modo, la estimación combina la probabilidad de pertenecer a cada grupo con la severidad media observada en cada uno de ellos.

Ambos modelos comparten dos limitaciones importantes. Por un lado, solo pueden entrenarse con observaciones completas, descartando los casos cuyo proceso de recuperación aún no ha concluido. Por otro, asumen una relación lineal entre las variables y la LGD, lo que limita su capacidad para capturar patrones más complejos. Estas dos limitaciones son el punto de partida que motiva los enfoques que veremos a continuación.

2.2 Análisis de Supervivencia

El análisis de supervivencia, que a partir de ahora abreviaremos como SA (Survival Analysis) es el conjunto de técnicas estadísticas que modelan la variable "tiempo hasta la ocurrencia de un evento". En medicina, donde más comúnmente es aplicado, este evento suele ser la muerte de un paciente a lo largo del estudio, donde nace el nombre de Análisis de Supervivencia. Sin embargo, su aplicación se extiende a cualquier campo donde se quiera modelar el tiempo hasta que ocurre un evento de interés,como el fallo de una componente en ingeniería, la reincidencia de un delito en criminología, o, como veremos en este tabajo, la recuperación de un importe en un proceso de recuperación de créditos en mora.

Una de las principales ventajas que presenta el SA es la capacidad de trabajar con datos censurados. Siguiendo con las aplicaciones médicas (pues son el origen de este método), imaginemos un estudio en el que se realiza un seguimiento a un grupo de pacientes para estudiar el tiempo hasta su fallecimiento. Al acabar el estudio algunos habrán fallecido y conoceremos exactamente cuándo. Para los que no han fallecido, sabemos que siguen vivos aunque no cuando morirán. Y que no lo hayan hecho durante el estudio no significa que no vayan a hacerlo. Estas observaciones son censuradas, y el SA nos permite no descartarlas porque contienen información valiosa. En este estudio para el caso de la LGD, el papel del paciente lo desempeña el préstamo en default, y el evento de interés es la recuperación del saldo pendiente. Al finalizar el periodo de datos disponible, algunos préstamos habrán cerrado su proceso de recobro y conoceremos exactamente cuánto se recuperó y cuándo.

Otros, sin embargo, seguirán en mora sin que el proceso haya concluido — son los episodios censurados de nuestro estudio. De ellos sabemos cuánto se ha recuperado hasta el momento del corte de datos, pero no la pérdida final que tendrán.

2.2.1 Función de supervivencia y tasa de riesgo

En el análisis de supervivencia se define T como la variable aleatoria que representa el tiempo hasta el evento de interés, y con ella $f(t)$ su función de densidad para $t \geq 0$ y $F(t)$ su función de distribución acumulada. Esta última por tanto recoge la probabilidad de que el evento haya ocurrido antes del instante t, es decir, $F(t) = P(T \leq t)$.

A partir de esta función se construyen dos conceptos fundamentales del SA. En primer lugar, la función de supervivencia $S(t)$ recoge lo contrario que $F(t)$, la probabilidad de que el evento no haya ocurrido hasta el momento t:

$$S(t) = P(T > t) = 1 - F(t) \tag{2.4}$$

Por construcción, $S(t)$ es una función decreciente que parte de $S(0) = 1$ y tiende a cero conforme t aumenta. Junto a ella, el segundo concepto fundamental es el de tasa de riesgo o *hazard rate*, definida como:

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{2.5}$$

La tasa de riesgo responde a la siguiente pregunta: dado que el evento no ha ocurrido hasta el momento t, ¿a qué velocidad ocurre justo en ese instante? Es decir, mide la velocidad instantánea con la que ocurre el evento en el instante t, condicionada a que todavía no haya ocurrido hasta ese momento.

Es útil definir también la función de riesgo acumulado, que agrega la tasa de riesgo instantánea desde el inicio del periodo hasta el momento t:

$$\Lambda(t) = \int_0^t \lambda(s)\, ds \tag{2.6}$$

Esta función permite relacionar directamente la tasa de riesgo acumulada con la función de supervivencia, de forma que:

$$S(t) = e^{-\Lambda(t)} \tag{2.7}$$

Aplicando estos conceptos al proceso de recuperación de créditos, $F(t)$ corresponde a la tasa de recuperación esperada hasta el momento t, mientras que $S(t)$ representa la pérdida esperada si el proceso de recuperación se detuviese en ese instante. Por su parte, la tasa de riesgo $\lambda(t)$ corresponde a la velocidad de recuperación medida sobre el saldo pendiente en el momento t tras el default.

2.2.2 El estimador de Kaplan-Meier

El estimador de Kaplan-Meier es la herramienta no paramétrica más utilizada para estimar la función de supervivencia a partir de datos observados [3]. Al ser no paramétrica, no asume ninguna distribución concreta para S(t), sino que esta se construye directamente a partir de los datos. La idea es sencilla: en cada momento que ocurre un evento, el estimador calcula la proporcion de individuos que siguen en riesgo y actualiza la estimación de la supervivencia de forma acumulada. Formalmente:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{2.8}$$

Donde t_i son los instantes en que ocurre algún evento, d_i es el número de eventos observados en ese instante y n_i es el número de individuos en riesgo justo antes de t_i.

El resultado es una función escalonada que decrece en cada instante en que se produce un evento y permanece constante entre eventos. En el contexto de este estudio, aplicado a los datos de recuperación, la curva de Kaplan-Meier proporciona una estimación de la proporción del saldo pendiente que no ha sido recuperada hasta cada momento t, sin necesidad de asumir ningún modelo concreto sobre cómo evoluciona dicho proceso.

Para comparar las curvas de supervivencia de dos o más grupos, el test estadístico más utilizado es el test log-rank. Este contraste no paramétrico evalúa la hipótesis nula de que las curvas de supervivencia de los grupos comparados son idénticas. El estadístico se construye comparando, en cada instante en que ocurre un evento, el número de eventos observados en cada grupo con el número esperado bajo la hipótesis nula, y agregando estas diferencias a lo largo de todo el periodo de seguimiento. Un p-valor inferior al nivel de significancia establecido indica que al menos dos de las curvas son estadísticamente distintas. En el contexto de este trabajo, dicho test se utilizará para contrastar si el proceso de recuperación difiere significativamente entre episodios con distinto tipo de terminación.

2.2.3 El modelo de Cox

El estimador de Kaplan-Meier nos aporta una primera visión de cómo evoluciona la recuperación a lo largo del tiempo, pero presenta una gran limitación: solo tiene en cuenta la dimensión temporal sin incorporar variables explicativas, tratando a todos los préstamos de la misma forma independientemente de sus características. Para incorporar esta información es necesario utilizar un enfoque distinto, que es precisamente lo que propone el modelo de Cox.

El modelo de riesgos proporcionales de Cox es uno de los enfoques más utilizados dentro del análisis de supervivencia [4]. Su principal atractivo reside en que permite estimar probabilidades de supervivencia para un préstamo concreto incorporando covariables explicativas, sin asumir una distribución específica para la tasa de riesgo base. Por ello se considera un modelo semi-paramétrico. La idea central es expresar la tasa de riesgo de un préstamo con características x como el producto de dos componentes: una tasa de riesgo base y un factor que recoge el efecto de dichas características:

$$\lambda(t \mid x) = \lambda_0(t) \exp(x'\beta) \tag{2.9}$$

Donde $\lambda_0(t)$ representa la tasa de riesgo base, común a todos los préstamos, y $\exp(x'\beta)$ es un multiplicador individual que escala la tasa base hacia arriba o hacia abajo en función de las características de cada individuo.

Los coeficientes β se estiman mediante verosimilitud parcial. En cada instante en que ocurre un evento, el modelo calcula la probabilidad de que sea precisamente ese individuo quien lo experimenta, frente al conjunto de individuos que siguen en riesgo en ese momento:

$$L_i = \frac{\exp(x_i'\beta)}{\sum_{j \in A_i} \exp\left(x_j'\beta\right)} \tag{2.10}$$

Donde x_i son las covariables del préstamo i que experimenta el evento, t_i es el instante en que dicho evento ocurre, y A_i es el conjunto de préstamos en riesgo justo antes de t_i. Es decir, se calcula la probabilidad de que sea precisamente el préstamo i el que experimente el evento en t_i, frente a todos los préstamos que siguen en riesgo en ese momento. Los coeficientes β se obtienen maximizando la suma de los logaritmos de estas probabilidades, sin necesidad de estimar directamente $\lambda_0(t)$.

Una vez obtenidos los coeficientes β, la función de supervivencia para un individuo con covariables x se expresa como:

$$S(t \mid x) = S_0(t)^{\exp(x'\beta)} \tag{2.11}$$

Donde $S_0(t)$ es la función de supervivencia base correspondiente a $\lambda_0(t)$, que se estima de forma no paramétrica una vez fijados los coeficientes β.

Es importante destacar que el modelo asume que el efecto de las covariables sobre la tasa de riesgo es multiplicativo y constante a lo largo del tiempo, razón por la que este recibe el nombre de modelo de riesgos proporcionales de Cox. Esta hipótesis de proporcionalidad temporal facilita la interpretación de los coeficientes, ya que permite analizar cómo cambia el riesgo relativo asociado a cada variable.

No obstante, el modelo de Cox también presenta limitaciones. La hipótesis de proporcionalidad suele no cumplirse en todos los casos y las relaciones entre covariables y riesgo pueden ser no lineales o incluir interacciones complejas. Estas restricciones motivan el uso de modelos más flexibles basados en aprendizaje automático.

2.2.4 Aplicación al modelado de la LGD

La aplicación del análisis de supervivencia a la estimación de la LGD parte de una idea sencilla: el proceso de recuperación de un crédito en default es, por naturaleza, un proceso temporal. Desde el momento en que se produce el default hasta que finaliza el proceso de recobro, puede transcurrir un tiempo considerable, durante el cual el banco va recuperando

fracciones del saldo pendiente. Esta dimensión temporal es precisamente la que el SA está diseñado para capturar.

En este estudio, el evento de interés es la recuperación del saldo pendiente del préstamo, y la función de supervivencia $S(t \mid x)$ recoge la proporción del saldo que el modelo estima que no habrá sido recuperada hasta el momento t, dadas las características del préstamo recogidas en x. De esta manera, la LGD estimada en el momento del default puede expresarse como:

$$LGD = S(K \mid x) \tag{2.12}$$

Donde K es el horizonte máximo de recuperación considerado, también conocido en la práctica bancaria como *Time-to-Workout*. Este horizonte delimita el periodo desde el default durante el cual se espera observar la mayor parte de las recuperaciones. A partir de K no se esperan recuperaciones sdicionales significativas y por tanto, el proceso de recobro se considera cerrado a efectos de la estimación. En lugar de fijar este horizonte de antemano mediante un criterio descriptivo o regulatorio, en este trabajo K se trata como un parámetro a seleccionar empíricamente: se elige el valor que ofrece mejor capacidad predictiva sobre los datos disponibles. El procedimiento concreto de selección se desarrolla en el Capítulo 5.

Siguiendo el hilo de esta idea, en este trabajo se propone estimar la pérdida esperada para préstamos que ya llevan un tiempo p en mora. Esta extensión, que forma parte de una de las contribuciones originales del trabajo, aprovecha la naturaleza condicional de la función de supervivencia para actualizar la estimación de la LGD en cualquier momento del *workout period*:

$$LGD_{\text{residual}}(p) = \frac{S(p + K \mid x)}{S(p \mid x)} \tag{2.13}$$

De este modo, la estimación de la LGD puede actualizarse dinámicamente conforme se dispone de nueva información sobre el proceso de recobro. En el capítulo 5 se detalla la adaptación de esta expresión y de las técnicas explicadas a los datos de recuperación crediticia del estudio.

2.3 Introducción al aprendizaje automático: árboles de decisión y *Random Forest*

2.3.1 Árboles de decisión

Un árbol de decisión es un modelo predictivo que divide el espacio de variables explicativas en regiones cada vez más homogéneas mediante la realización de una serie de preguntas o divisiones binarias sucesivas. La construcción del árbol comienza en el nodo raíz, que contiene todas las observaciones, y se va dividiendo de forma recursiva hasta alcanzar un criterio de parada.

En cada nodo se busca la variable y el punto de corte que mejor separan las observaciones según la variable objetivo [5]. En problemas de regresión, como es el caso de la predicción de la LGD, el criterio habitual es minimizar la suma de cuadrados de los residuos (RSS) dentro de cada grupo resultante de la división:

$$RSS = \sum_{i \in R_1} (y_i - \bar{y}_{R_1})^2 + \sum_{i \in R_2} (y_i - \bar{y}_{R_2})^2 \tag{2.14}$$

Donde R_1 y R_2 son las dos regiones resultantes del *split*, y $\bar{y}_{R_1}$ y $\bar{y}_{R_2}$ son sus respectivas medias. El algoritmo evalúa todos los posibles puntos de corte de todas las variables y selecciona aquel que produce la mayor reducción del RSS.

Una vez construido el árbol, la predicción para una nueva observación se obtiene recorriendo el árbol desde la raíz hasta un nodo terminal, siguiendo las divisiones que correspondan a los valores de sus variables explicativas. La predicción final es la media de la variable objetivo de todas las observaciones de entrenamiento que cayeron en ese nodo terminal.

La principal ventaja de los árboles de decisión es su capacidad para capturar relaciones no lineales e interacciones entre variables sin necesidad de especificarlas de antemano. Sin embargo, presentan una limitación importante: tienden a sobreajustarse a los datos de entrenamiento, generando árboles profundos que capturan el ruido de la muestra y pierden capacidad de generalización. Esta limitación motiva el desarrollo de métodos de *ensemble* como el *Random Forest*.

2.3.2 Extensión a *Random Forests*

El *Random Forest* es un método de *ensemble* que combina un gran número de árboles de decisión independientes con el objetivo de reducir la varianza de la predicción [6]. Un árbol individual tiene alta varianza, ya que es muy sensible a los datos concretos con los que fue entrenado, de forma que pequeños cambios en la muestra pueden producir árboles muy distintos. Al promediar las predicciones de muchos árboles construidos sobre muestras y subconjuntos de variables diferentes, esa variabilidad individual tiende a cancelarse y el resultado conjunto es mucho más estable que cualquier árbol por separado. Además, estos modelos suelen ofrecer medidas de importancia de variables, útiles para interpretar qué factores contribuyen más a la predicción.

La diversidad entre árboles se consigue mediante dos mecanismos principales. El primero es el *bagging* o *bootstrap aggregation*: cada árbol se entrena sobre una muestra aleatoria con reemplazamiento del conjunto de entrenamiento, de forma que cada árbol observa una versión ligeramente distinta de los datos. El segundo es la aleatoriedad en la selección de variables: al hacer una división en cada nodo del árbol, no se busca el mejor corte entre todas las variables disponibles, sino solo entre un subconjunto aleatorio de k variables. Esto evita que todos los árboles utilicen siempre las mismas variables principales y favorece que sean suficientemente distintos entre sí.

En problemas de regresión, como la predicción de la LGD, la predicción final del *Random Forest* para una nueva observación es el promedio de las predicciones de todos los árboles individuales:

$$\hat{f}_{RF}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x) \tag{2.15}$$

Donde B es el número total de árboles del bosque y $\hat{f}_b(x)$ es la predicción realizada por el árbol b para una observación con características x. De este modo, el modelo mantiene la capacidad de los árboles para adaptarse a patrones complejos, pero obtiene resultados más estables y menos dependientes del ruido de los datos de entrenamiento, reduciendo así el sobreajuste. En el apartado siguiente se presenta su extensión al análisis de supervivencia, que permite aplicar estas ventajas al modelado de tiempos hasta evento con observaciones censuradas

2.4 La integración de ambos enfoques: *Random Survival Forests*

Los *Random Survival Forests* (RSF) combinan la lógica de los *Random Forests* con los principios del análisis de supervivencia, adaptando cada árbol se al contexto de supervivencia [7]. Un árbol de supervivencia en lugar de predecir un valor continuo, estima una función de supervivencia en sus nodos terminales. Por ello, las divisiones ya no se eligen minimizando el RSS, sino mediante el estadístico *log-rank*, que mide la diferencia entre las curvas de supervivencia de los dos grupos resultantes del *split*. El mejor *split* es aquel que maximiza esa diferencia, es decir, el que mejor separa a los individuos según su velocidad de recuperación. En cada nodo terminal, la función de supervivencia se estima mediante Nelson-Aalen, lo que permite utilizar tanto las observaciones completas como las censuradas hasta el momento en que se dispone de información [8, 9].

El objetivo final es obtener una función de supervivencia individual $S(t \mid x)$ para cada observación, es decir, la probabilidad de que el evento no haya ocurrido hasta el momento t dado el perfil de características x. Para ello, se construyen B árboles d supervivencia independientes mediante *bagging* y selección aleatoria de variables, y se promedian sus predicciones :

$$\hat{S}(t \mid x) = \frac{1}{B}\sum_{b=1}^{B}\hat{S}_b(t \mid x) \tag{2.16}$$

Donde $\hat{S}_b(t \mid x)$ es la función de supervivencia estimada por cada árbol b para una observación de características x.

En este trabajo, los RSF se plantean como una alternativa para complementar los enfoques tradicionales y los modelos de supervivencia clásicos. Su aplicación permitirá analizar si la combinación de censura, tiempo e interacciones no lineales mejora la capacidad predictiva en la estimación de la LGD.

Capítulo 3

Datos y construcción de la muestra

En este capítulo se describirá la base de datos empleada en el estudio, así como los criterios aplicados para transformar la información original en la muestra final de modelización. En particular, se detallarán las fuentes disponibles, el periodo temporal cubierto, el tipo de operaciones incluidas y las principales restricciones utilizadas para garantizar la consistencia de los datos.

3.1 Fuente de datos

El dataset utilizado es el *Single-Family Loan-Level Dataset* publicado por Freddie Mac, una de las fuentes públicas más completas sobre hipotecas residenciales en Estados Unidos. Su publicación responde a un mandato de transparencia de la Federal Housing Finance Agency (FHFA) orientado a facilitar el desarrollo de modelos de riesgo de crédito [10]. La información se distribuye en dos ficheros: uno de originación, con las características estáticas del préstamo en el momento de concesión, y otro de *performance*, con observaciones mensuales como el saldo, mora, recuperaciones y terminación del préstamo.

Por limitaciones computacionales, se utiliza la muestra aleatoria que Freddie Mac publica junto al dataset completo, formada por 50.000 préstamos por año de originación. Se seleccionan las datos de 2005 a 2025, cubriendo la crisis financiera global de 2008, el ciclo de recuperación posterior, la pandemia de Covid-19 y su etapa posterior. Dado que el objetivo del estudio es estimar la pérdida en caso de impago, se retendrán solo los préstamos que entran en default en algún momento de su vida. La muestra inicial contiene 53.191 préstamos con algún episodio de impago y 4.733.126 observaciones mensuales. No obstante, tras aplicar los criterios de depuración descritos en los apartados posteriores, se trabajará con una muestra de menor tamaño.

La definición completa de todas las variables del dataset puede consultarse en Freddie Mac [10]. En el Anexo A se recogen únicamente las variables construidas durante la preparación de datos y las variables de originación y performance utilizadas en el análisis, con una descripción del tratamiento aplicado en cada caso.

3.2 Construcción de la muestra y episodios de default

3.2.1 Filtrado inicial de los datos de *performance*

La definición de default constituye el punto de partida del análisis, ya que fija el momento a partir del cual comienza a observarse el proceso de recuperación. En la práctica regulatoria, la identificación del default puede depender de distintos criterios y casuísticas. En este trabajo se adopta una versión simplificada basada en el umbral de 90 días de mora: un préstamo se considera en default en el primer mes en que acumula 90 o más días de atraso. Esta decisión se toma por motivos de simplicidad, pues el objetivo principal del estudio es explorar la aplicación de nuevas técnicas de modelización de la LGD más que construir un modelo plenamente regulatorio.

Para identificar dicho momento se utiliza la variable `current_loan_delinquency_status` del fichero de *performance* de Freddie Mac. Esta variable recoge, para cada mes de vida del préstamo, el grado de atraso del prestatario según el método de la Mortgage Bankers Association (MBA): toma el valor 0 cuando el préstamo está al corriente, 1 cuando acumula entre 30 y 59 días de retraso, 2 entre 60 y 89 días, y así sucesivamente. Por tanto, en este estudio se considera que un préstamo entra en default en el primer mes en que `current_loan_delinquency_status` ≥ 3. Dado que el objetivo del trabajo es estimar la LGD, la muestra se restringe a los préstamos que alcanzan este umbral en algún momento de su vida, excluyendo aquellos que nunca llegan a estar en default. A partir de la primera observación con `current_loan_delinquency_status` ≥ 3 comienza el periodo de recobro, durante el cual se registran los flujos de caja necesarios para el cálculo de la LGD.

Además, en esta etapa se calculan dos variables que se utilizarán a lo largo del análisis: `months_since_default`, que mide el número de meses transcurridos desde la entrada en default, y la EAD (*Exposure at Default*), definida como el saldo vivo del préstamo en el primer mes de default. Ambas variables se documentan con mayor detalle en el Anexo A.

3.2.2 Criterios de exclusión

Tras restringir la muestra a los préstamos que entran en default, se aplican distintos filtros adicionales para excluir operaciones que no resultan adecuadas para el cálculo de la LGD bajo el enfoque de *workout period*.

El campo `zero_balance_code` es una variable categórica que identifica el motivo por el que el saldo del préstamo se reduce a cero y, por tanto, el evento que pone fin a su ciclo de vida. Permite distinguir entre distintos tipos de terminación, como el prepago voluntario (código 01), la disposición REO[1] (código 09), las ventas cortas, las ventas a terceros o las retiradas administrativas.

Una vez definida esta variable se eliminan los 1.311 préstamos con 96 como `zero_balance_code`, correspondientes a operaciones retiradas del dataset por un defecto administrativo previo a cualquier evento de terminación económica real. Su inclusión no aportaría información

[1]REO son las siglas de *Real Estate Owned*. Se refiere a viviendas que pasan a ser propiedad del acreedor, normalmente después de una ejecución, hasta que se venden.

sobre el proceso de recuperación y únicamente introduciría ruido en el análisis.

En segundo lugar, se excluyen los préstamos que han sido objeto de reestructuraciones contractuales (`modification_flag = Y`) o que presentan un plan de pago diferido activo. Esta decisión no pretende reproducir el tratamiento de un modelo regulatorio, donde este tipo de operaciones sí debería analizarse de forma específica, sino simplificar la muestra y mantener el foco del trabajo en la aplicación del análisis de supervivencia a episodios de recuperación comparables. En estos casos, la trayectoria de recuperación puede verse alterada por una intervención del acreedor, por lo que incorporarlos requeriría un tratamiento adicional que queda fuera del alcance de este estudio. Al aplicar este filtro, se eliminan 24.151 préstamos adicionales, quedando la muestra en 27.729 préstamos.

En tercer lugar, se eliminan los préstamos con $EAD = 0$ en el momento del default, ya que una exposición nula impide calcular una pérdida económicamente interpretable. Tras este filtro, la muestra queda en 27.086 préstamos únicos.

Por último, se eliminan los episodios en estado REO para los que no se dispone de código de cierre ni de saldo vivo igual a cero. Estos 59 episodios corresponden a préstamos cuya propiedad ha sido adjudicada por el banco, pero cuyo proceso de venta no se ha completado al final del periodo de observación, por lo que no se dispone de información sobre las recuperaciones finales. Una alternativa para tratarlos sería aplicar un modelo de *haircut* que estimase el valor de recuperación esperado a partir del saldo pendiente y de las características del inmueble, pero su desarrollo queda fuera del alcance de este trabajo. Tras aplicar todos los filtros descritos, la muestra final queda formada por 27.027 préstamos únicos.

3.2.3 Definición de episodio de default

En este trabajo, la unidad fundamental de análisis no es el préstamo sino el episodio de default: un ciclo continuo de morosidad que comienza cuando el préstamo entra en mora y termina cuando se produce una resolución definitiva o se agota el periodo de observación. La razón de trabajar a nivel de episodio es que un mismo préstamo puede impagar, recuperarse y volver a impagar pasado un tiempo considerableen más de una ocasión. Tratar ambos ciclos como un único episodio mezclaría situaciones de riesgo estructuralmente distintas.

Para distinguir episodios de default es necesario definir previamente cuándo se considera que un préstamo ha curado. En este trabajo se adopta un criterio sencillo: se construye la variable `is_cured_t`, que toma valor 1 cuando el préstamo acumula al menos tres meses consecutivos con `dq_num` < 3 y no se encuentra en estado REO. Este criterio permite identificar periodos en los que el préstamo parece haber salido de la situación de default, sin pretender reproducir de forma completaa una definición regulatoria de curación.

Una vez definida la curación, se establece el umbral de separación entre episodios: si un préstamo permanece en estado curado durante al menos 9 meses consecutivos antes de volver a impagar, el nuevo ciclo de morosidad se trata como un episodio independiente. Si reincide antes de ese umbral, ambos momentos de default y los meses intermedios se consolidan dentro del mismo episodio.

De esta manera, cada episodio independiente se identifica mediante la variable `loan_id`, construida a partir de la concatenación del identificador original del préstamo y del número de episodio. Los meses de curación que separan episodios distintos se eliminan del panel, al corresponder a periodos de no-default que no forman parte del proceso de resolución.

3.2.4 Resultado de la segmentación

La segmentación descrita da lugar a 28.720 episodios únicos, correspondientes a los 27.027 préstamos de la muestra. De ellos, 1.469 préstamos (5,4 %) presentan más de un episodio de default, lo que confirma la existencia de un grupo relevante de prestatarios reincidentes.

Cada episodio se clasifica según su estado al final del periodo de observación. Los episodios completos (27.181) son aquellos que alcanzan una resolución definitiva, ya sea mediante un evento de terminación o mediante una curación confirmada, y para los que se dispone de una LGD observada. Por el contrario, los episodios abiertos (1.539) son aquellos para los que no se observa una resolución dentro del periodo de análisis. Estos casos constituyen observaciones censuradas por la derecha en el modelo de supervivencia, ya que se sabe que el proceso de recobro no ha concluido, pero no se observa cuándo ni cómo finalizará.

Dentro de los episodios completos, la variable `terminacion_cura` distingue entre terminaciones favorables (`terminacion_cura = 1`) y desfavorables (`terminacion_cura = 0`). Las primeras incluyen cancelaciones voluntarias, ventas como préstamos *reperforming* o curaciones definitivas, y representan el 48,2 % de los episodios completos (13.112). Las segundas incluyen adjudicaciones REO, ventas cortas, *charge-offs* u otros eventos de pérdida, y representan el 51,8 % restante (14.069).

Finalmente, la variable `duration_months` recoge el número de meses transcurridos desde el primer mes en default hasta el último mes observado del episodio, y constituye la variable de tiempo utilizada en los modelos de supervivencia desarrollados en el Capítulo 5. En los episodios abiertos, esta variable mide el tiempo observado hasta el final de la ventana de seguimiento.

3.3 Construcción de la variable objetivo

En este trabajo se calcula la LGD no solo en el momento del default, sino también en cada mes t del episodio. El valor en $t = 0$, denotado como $LGD_{i,0}$, constituye la LGD definitiva del episodio y es la variable objetivo de los modelos estáticos desarrollados en el Capítulo 5. El valor en momentos posteriores, $LGD_{i,t}$, mide la pérdida esperada restante a partir del saldo vivo y de los flujos de caja pendientes desde ese mes hasta la resolución del episodio, y será la base de la predicción dinámica de la LGD residual descrita en el apartado 5.5. Cabe señalar que esta variable no está disponible directamente en el dataset de Freddie Mac, sino que ha sido construida a partir de los flujos de caja observados mediante la siguiente expresión:

$$LGD_{i,t} = 1 - \frac{1}{UPB_{i,t}} \sum_{s=t}^{T} \frac{CF_{i,s}}{(1 + r_i)^{(s-t)/12}} \tag{3.1}$$

Donde $UPB_{i,t}$ es el saldo vivo del episodio i en el momento t, $CF_{i,s}$ representa los flujos de caja recuperados en cada periodo posterior s - su construcción se detalla en el siguiente apartado- y r_i es la tasa de descuento asociada al episodio. El descuento se aplica de forma relativa al momento t en que se evalúa la LGD parcial, y no al inicio del episodio. De este modo, $LGD_{i,t}$ representa la pérdida residual estimada desde el mes t, mientras que $LGD_{i,0}$, calculada en el primer mes de default, corresponde a la LGD definitiva del episodio completo y constituye la variable objetivo del estudio.

3.3.1 Construcción de los flujos de caja

La construcción de los flujos de caja, recogidos en la variable `cash_flow_t`, distingue varios casos según el estado del préstamo en cada mes del episodio. En primer lugar, en el mes 0 el flujo de caja se fija igual a cero por convención, ya que como se ha comprobado el mes de entrada en default no genera recuperación.

En los meses intermedios sin evento de terminación, el flujo se aproxima mediante la reducción del saldo vivo entre dos meses consecutivos:

$$CF_t = UPB_{t-1} - UPB_t \tag{3.2}$$

Esta diferencia refleja los pagos parciales que el prestatario pueda realizar durante el proceso de mora.

Cuando el episodio termina por prepago, identificado mediante `zero_balance_code = 01`, el flujo de caja del último mes de vida de dicho préstamo se iguala al campo `zero_balance_removal_upb`, que recoge el saldo total pendiente en el momento del cierre. Este es el único campo informado para dicho código en el dataset de Freddie Mac.

En los meses de terminación asociados a cualquier otro código de cierre, el flujo recoge las recuperaciones netas de gastos e intereses acumulados impagados:

$$\begin{aligned} CF = {} & \texttt{net\_sale\_proceeds} + \texttt{mi\_recoveries} + \texttt{non\_mi\_recoveries} \\ & - \texttt{total\_expenses} - \texttt{delinquent\_accrued\_interest} \end{aligned} \tag{3.3}$$

Los gastos totales, `total_expenses`, se calculan como el máximo entre el campo agregado `expenses` y la suma de sus cuatro componentes individuales: costes legales, mantenimiento, impuestos y seguros, y gastos varios. Esta decisión se debe a que el campo agregado solo estaba informado en 54 casos, frente a 2.433 observaciones con componentes individuales disponibles.

Finalmente, para los episodios que terminan en curación definitiva sin código de cierre —es decir, préstamos que salen de la situación de default porque el prestatario se pone al día en sus pagos—, se genera un flujo de caja artificial en el último mes:

$$CF = \texttt{current\_actual\_upb} \tag{3.4}$$

Con ello se asume la recuperación total del saldo pendiente en el momento en que el prestatario retoma el pago regular del préstamo.

3.3.2 Tasa de descuento

La tasa de descuento de cada episodio se fija como el `current_interest_rate` observado en el primer mes del episodio, expresado en tanto por uno. Se toma este criterio por simplicidad y coherencia con la información disponible en el dataset, aunque en la práctica la tasa dependería del uso que vaya a tener el modelo, ya sea regulatorio, contable o interno. Esta decisión puede hacer que las LGD estimadas sean más altas o más bajas en conjunto, pero no afecta a la segmentación y a la comparación entre modelos, ya que se aplica la misma tasa a toda la muestra. Por último, aunque el nombre de la variable podría sugerir variación temporal, en la práctica permanece constante durante la vida del préstamo, dado que el dataset empleado recoge exclusivamente hipotecas a tipo fijo.

3.3.3 Tratamiento de episodios curados y censurados

Los 8.427 episodios que terminan en curación definitiva sin código de cierre reciben el flujo de caja artificial descrito anteriormente, asumiendo recuperación total del saldo pendiente en el momento en que el prestatario retoma el pago regular. Esta aproximación es coherente con la idea de curación: si el prestatario vuelve a atender el préstamo con normalidad, se considera que el saldo pendiente puede recuperarse íntegramente.

Para los 1.479 episodios abiertos se calcula una LGD aproximada aplicando la misma fórmula con los flujos disponibles hasta el último mes observado y asumiendo que el saldo pendiente no se recupera. Esta estimación debe interpretarse con cautela, ya que no representa la pérdida real del episodio, sino una aproximación parcial útil únicamente con fines descriptivos. Al no observarse la resolución final, estos episodios no se utilizarán para evaluar la capacidad predictiva del modelo en validación. No obstante, como se verá en el Capítulo 4, el análisis de supervivencia permite incorporarlos al entrenamiento como observaciones censuradas por la derecha, aprovechando la información registrada hasta el final de la ventana de observación.

Tras calcular la LGD para todos los episodios, se eliminan aquellos que presentan valores fuera del rango $[-0,5; 1,5]$ en algún mes del episodio. Este filtro actúa como un criterio de depuración de valores extremos u *outliers* de la variable objetivo: afecta a 218 episodios y permite descartar observaciones con inconsistencias en los datos de recuperación que distorsionan el análisis posterior. Tras esta depuración, la muestra final queda formada por 28.502 episodios de default.

3.4 Análisis descriptivo de la muestra

La muestra final utilizada en este trabajo comprende 28.502 episodios de default correspondientes a 26.872 préstamos únicos, con un total de 441.968 observaciones mensuales. El 94,7 % de los préstamos presenta un único episodio de default, mientras que el 5,3 % restante reincide en mora en al menos una ocasión.

En cuanto al tipo de resolución, el 94,8 % de los episodios cuenta con una resolución definitiva observada, mientras que el 5,2 % permanece abierto al final del periodo de observación. Dentro de la muestra total, las terminaciones desfavorables representan el 49,2 %, las favorables el 45,6 % y los episodios abiertos el 5,2 % restante. La Tabla 3.1 recoge esta distribución.

Cuadro 3.1: Distribución de episodios por tipo de terminación

Tipo de terminación	N	%
Desfavorable	14.024	49,2
Favorable	12.999	45,6
Abierto	1.479	5,2
Total	28.502	100,0

Entre los eventos de terminación observados, el más frecuente es la disposición REO (`zero_balance_code = 09`), presente en el 41,5 % de los episodios con terminación. Le siguen el prepago voluntario (`zero_balance_code = 01`, 24,7 %), la venta corta o *charge-off* (`zero_balance_code = 03`, 18,3 %) y la venta a terceros (`zero_balance_code = 02`, 11,9 %).

3.4.1 Distribución de la LGD

La LGD definitiva de cada episodio se obtiene como el valor de LGD_t en $t = 0$, es decir, el valor calculado en el primer mes del episodio considerando todos los flujos de caja futuros descontados. Los estadísticos principales se recogen en la Tabla 3.2.

Cuadro 3.2: Estadísticos descriptivos de la LGD

Estadístico	Valor
Media	0,307
Mediana	0,118
Desviación típica	0,346
Mínimo	-0,414
Percentil 25	0,022
Percentil 75	0,556
Máximo	1,377

La distribución de la LGD presenta una marcada asimetría positiva y una forma bimodal característica de las carteras hipotecarias. Una parte importante de los episodios se

concentra en valores próximos a cero: el 48,7 % presenta una LGD inferior a 0,10, principalmente en terminaciones favorables. Al mismo tiempo, existe otra concentración relevante en valores cercanos a la unidad, con un 11,3 % de episodios por encima de 0,90. Esta bimodalidad refleja la naturaleza del proceso de resolución hipotecaria: los préstamos que se curan o se cancelan voluntariamente generan pérdidas reducidas, mientras que los que terminan en adjudicación o venta forzosa concentran la mayor parte de las pérdidas. Los valores negativos de LGD, posibles cuando las recuperaciones superan el saldo expuesto, representan una fracción marginal de la muestra.

3.4.2 LGD por tipo de terminación

Como cabe esperar por la propia construcción de la variable objetivo, la heterogeneidad de la LGD entre grupos es pronunciada, como se observa en la Tabla 3.3 y en la Figura 3.1. Los episodios con terminación favorable presentan una LGD media de 0,037 y una mediana de 0,022: al recuperarse prácticamente todo el capital expuesto, su distribución se concentra cerca de cero. Los episodios desfavorables, con una LGD media de 0,486 y una mediana de 0,474, presentan en cambio una distribución mucho más dispersa a lo largo de todo el rango, lo que refleja la variabilidad de las pérdidas en los procesos con este tipo de terminación. Por último, los episodios abiertos, cuya LGD aproximada asume pérdida total del saldo pendiente, se agrupan en la cola derecha con valores próximos a 1.

Cuadro 3.3: LGD por tipo de terminación

Tipo de terminación	N	LGD media	LGD mediana	Desv. típica
Favorable	12.999	0,037	0,022	0,049
Desfavorable	14.024	0,486	0,474	0,304
Abierto	1.479	0,989	0,999	0,041
Total	28.502	0,307	0,118	0,346

3.4.3 Duración de los episodios

La distribución de la duración de los episodios presenta una marcada asimetría positiva, como se aprecia en la Figura 3.2. La mayoría de los episodios se resuelve en los primeros meses tras el default —la mediana se sitúa en los 10 meses—, pero existe una cola prolongada de casos que se extienden durante varios años, con un máximo de 179 meses en los episodios resueltos.

El tipo de resolución marca una gran diferencia en la velocidad del proceso. Los episodios favorables, que incluyen curaciones y prepagos, se resuelven con una mediana de 5 meses. Los desfavorables, en cambio, requieren una mediana de 16 meses y presentan una cola considerablemente más larga: el percentil 90 se sitúa en 42 meses, frente a los 18 meses de los favorables. Esto refleja que los procesos de ejecución hipotecaria y venta forzosa son más lentos y variables que una curación o cancelación voluntaria.

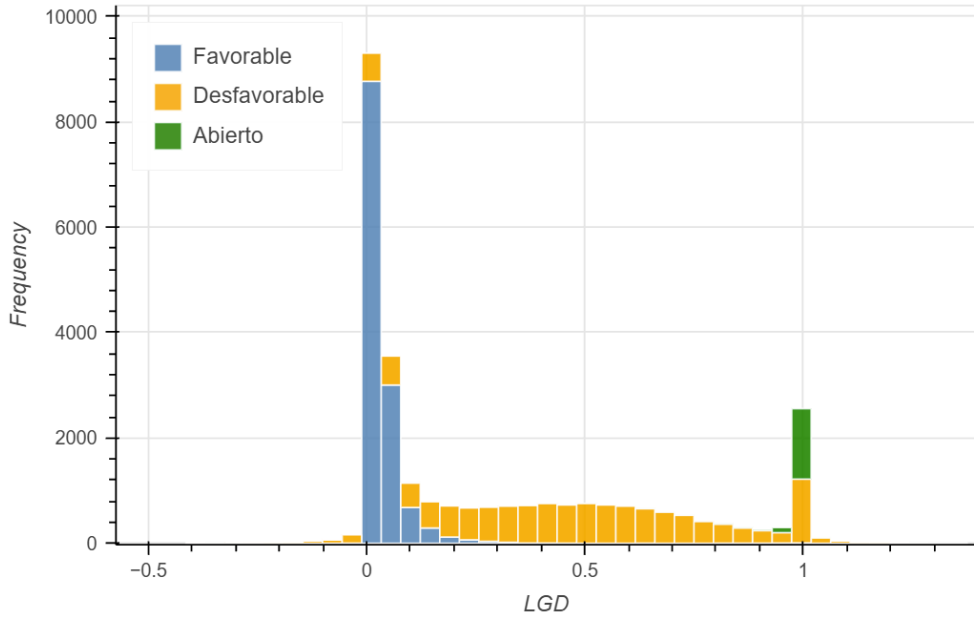

Figura 3.1: Distribución de la LGD según el tipo de terminación

Los episodios abiertos presentan una duración media similar a la de los favorables, aunque por razones completamente distintas: no se trata de episodios que se hayan resuelto rápido, sino de los más recientes, para las que la ventana de observación es más corta. En este grupo se incluyen 166 episodios con duración igual a cero, correspondientes a préstamos que acababan de entrar en default en el último periodo de observación.

Estas diferencias en la velocidad de resolución tienen implicaciones directas para la elección del horizonte K: un valor demasiado corto dejaría sin cubrir una parte relevante de los episodios desfavorables, mientras que uno excesivamente largo añadiría ruido sin mejorar necesariamente la estimación. Las líneas verticales de la Figura 3.2, correspondientes a los percentiles 50, 86, 88 y 90, anticipan algunos de los horizontes que se explorarán en el Capítulo 5 como candidatos para la definición de K.

3.5 Selección e ingeniería de variables

El proceso de selección parte de las 31 variables disponibles en el fichero de originación, a las que se añaden tres variables construidas a partir del fichero de *performance*: el tipo de interés vigente en el momento del default, `current_interest_rate_ep`; la antigüedad del préstamo en el momento del impago, `loan_age`, que recoge los meses transcurridos desde la originación hasta el default; y el número de episodio de default, `default_episode`, que indica si se trata del primer impago del prestatario o de uno posterior. En una primera etapa se excluyen las variables que no pueden utilizarse como predictores sin incurrir en problemas metodológicos: los identificadores del préstamo, las variables que describen el resultado del episodio (`resolucion`, `terminacion_cura` y `duration_months`), cuya inclusión constituiría *data leakage*, las fechas cuya información ya está recogida en otras variables, y las variables de vendedor y *servicer*, no disponibles de forma fiable

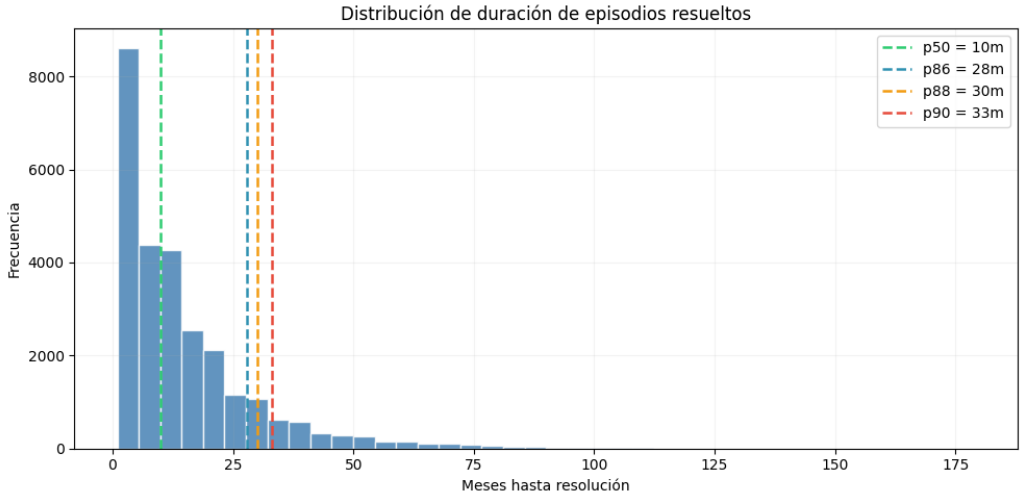

Figura 3.2: Distribución de la duración de episodios resueltos

en el momento del default. La localización geográfica se representa mediante la variable `postal_region`, construida a partir del primer dígito del código postal, que agrupa las observaciones en 9 regiones geográficas evitando el exceso de categorías que supondría trabajar directamente con los 50 estados o los códigos postales completos.

Sobre las 25 variables candidatas resultantes se aplican dos filtros de calidad: se eliminan las variables con más del 70 % de valores ausentes y aquellas que solo presentan una categoría o un valor constante en todo el conjunto de datos. Tras estos filtros quedan 22 variables candidatas. Los porcentajes de valores ausentes por variable se recogen en el Anexo A.

3.5.1 Análisis univariante

Se analiza la relación individual de cada variable con la LGD mediante tests no paramétricos. Para las variables numéricas se utiliza la correlación de Spearman, robusta frente a distribuciones no normales como la de la LGD. Para las variables categóricas se aplica el test de Kruskal-Wallis. Se conservan las variables con p-valor inferior a 0,01, criterio que todas las variables seleccionadas cumplen con claridad.

De las 12 variables numéricas, 9 resultan significativas, con `current_interest_rate_ep` e `interest_rate` como las más correlacionadas con la LGD ($\rho \approx 0,44$ en ambos casos), seguidas de `original_loan_term`, `loan_age`, `cltv`, `credit_score`, `ltv`, `dti` y `num_units`. Las tres variables descartadas, `mi_pct`, `num_borrowers` y `original_upb`, no muestran relación estadísticamente significativa. Las 10 variables categóricas candidatas resultan todas significativas según el test de Kruskal-Wallis, con `channel` y `postal_region` como las más discriminantes. `default_episode` también resulta significativa (estadístico = 134,3), aunque con una distribución muy concentrada en el primer episodio (94,3 % de los casos).

3.5.2 Análisis de multicolinealidad

El análisis de correlaciones entre variables numéricas y el cálculo del Factor de Inflación de la Varianza (VIF) revelan dos pares con multicolinealidad severa. El par formado por `current_interest_rate_ep` e `interest_rate` presenta una correlación de Spearman de 0,99 y un VIF superior a 38 en ambos casos; se conserva `current_interest_rate_ep` por ser el tipo vigente en el momento del default y, por tanto, más relevante para el proceso de recobro. El par formado por `cltv` y `ltv` presenta una correlación de 0,92 y un VIF próximo a 10. Aunque inicialmente se descarta `ltv` por tener menor correlación con la LGD, el análisis de Information Value posterior revela que discrimina mejor que `cltv`, por lo que se revierte la decisión y se conserva `ltv`. Tras este paso quedan 17 variables candidatas.

3.5.3 Selección multivariante por AIC

Se realiza una selección *stepwise* hacia adelante sobre una regresión lineal de la LGD, utilizando el Criterio de Información de Akaike (AIC) como criterio de selección. Dado que el modelo lineal requiere variables numéricas, las variables categóricas se transforman en variables *dummy* antes de su inclusión, tomando como categoría de referencia la más frecuente en cada caso. El AIC penaliza la complejidad del modelo de forma que una variable solo se incorpora si su aportación marginal supera el coste de incluirla. En este caso el algoritmo selecciona 24 variables. Sobre el modelo resultante se aplica la corrección de Bonferroni para controlar la tasa de error de tipo I en comparaciones múltiples. Los p-valores detallados y los resultados completos del modelo se recogen en el Anexo A.

3.5.4 Weight of Evidence e Information Value

Las variables numéricas se discretizan siguiendo una metodología basada en el Weight of Evidence (WOE) y el Information Value (IV), ampliamente utilizada en la construcción de modelos de *scoring* y LGD en el sector financiero. El WOE mide el poder discriminante de cada tramo de una variable respecto a la LGD, y el IV agrega dicho poder a nivel de variable. Los tramos de cada variable se obtienen a partir del análisis WOE/IV, buscando cortes que aumenten el Information Value y respeten cuatro criterios: un mínimo del 10 % de población por tramo, monotonía de la LGD media entre tramos, no solapamiento de los intervalos de confianza al 95 % entre tramos adyacentes, y entre 2 y 10 categorías por variable. Los resultados detallados del WOE e IV por variable y tramo se recogen en el Anexo A.

De los resultados del IV se concluye que `current_interest_rate_ep` es la variable con mayor poder predictivo, con un poder moderado, seguida de `loan_age` (IV = 0,046) y `ltv` (IV = 0,041), ambas con poder débil. El resto de variables numéricas presentan poder débil o nulo. Estos resultados, junto con la evidencia del análisis multivariante, llevan a descartar `dti` y `credit_score`, cuyo IV confirma que discrimina la probabilidad de default pero no la pérdida una vez producida.

3.5.5 Variables finales

La Tabla 3.4 recoge las variables seleccionadas para los modelos desarrollados en este estudio, junto con los rangos de discretización aplicados a las variables numéricas. Estos rangos proceden del análisis WOE/IV descrito anteriormente, y se han fijado buscando tramos con mayor capacidad discriminante y cumpliendo los criterios de monotonía y tamaño mínimo de población. La descripción de cada categoría, así como la distribución de observaciones por tramo o categoría, se detallan en el Anexo A.

Cuadro 3.4: Variables seleccionadas para el entrenamiento de los modelos

Variable	Tipo	Rangos o categorías
`current_interest_rate_ep`	Numérica	$< 4\,\%$, 4–5 %, 5–6 %, 6–7 %, $> 7\,\%$
`ltv`	Numérica	$< 60\,\%$, 60–80 %, 80–90 %, 90–100 %, $> 100\,\%$
`original_loan_term`	Numérica	≤ 15 años, 16–30 años, > 30 años
`loan_age`	Numérica	≤ 34 meses, 35–59 meses, 60–99 meses, > 99 meses
`loan_purpose`	Categórica	P, N, C
`occupancy_status`	Categórica	P, I, S
`property_type`	Categórica	SF, PU, CP, CO, MH
`first_time_homebuyer_flag`	Categórica	Y, N
`postal_region`	Categórica	1–9

Capítulo 4

Metodología

Este capítulo describe la metodología empleada para transformar la muestra construida en el Capítulo 3 en un dataset de supervivencia, entrenar los modelos predictivos y evaluar su capacidad para estimar la evolución temporal de la pérdida en caso de impago. El enfoque combina técnicas de análisis de supervivencia con procedimientos específicos de validación, ponderación y calibración, con el objetivo de aprovechar tanto los episodios completos como los episodios censurados.

La metodología se organiza en cinco bloques. En primer lugar, se presenta la construcción del dataset de supervivencia y el sistema de pesos aplicado a las observaciones. En segundo lugar, se define la estrategia de partición entre entrenamiento y prueba, prestando especial atención al tratamiento de los casos censurados. En tercer lugar, se describen las métricas de evaluación utilizadas para analizar discriminación y calibración. En cuarto lugar, se expone el procedimiento de optimización de hiperparámetros. Finalmente, se introduce la calibración *post-hoc* de los modelos, necesaria para ajustar las predicciones a los niveles observados de pérdida.

4.1 Construcción del Dataset de Supervivencia y Sistema de Pesos

El análisis de supervivencia requiere transformar el panel de *performance*, que contiene una fila por episodio y mes, en un formato en el que cada observación describa la evolución temporal de las recuperaciones. En este trabajo no se modela únicamente si el episodio se resuelve o no, sino la distribución mensual de los flujos recuperados en relación con la exposición inicial. Por ello, cada fila del dataset de supervivencia contiene tres elementos principales:

- `time`: número de meses transcurridos desde el inicio del episodio de default, medido mediante `months_since_default`.
- `event`: indicador binario que toma valor 1 si en ese mes se produce un flujo de caja positivo y 0 en caso contrario.

- `weight`: fracción del saldo expuesto recuperada en ese mes, definida como CF_t/EAD.

El uso de pesos es el elemento que conecta el análisis de supervivencia con la LGD. En lugar de asignar la misma importancia a todos los meses con evento, el modelo pondera cada recuperación por su peso económico relativo. Así, un flujo de caja elevado en relación con la EAD contribuye más al ajuste que una recuperación marginal.

Formalmente, cada observación mensual puede representarse como:

$$(t_{ij}, \delta_{ij}, w_{ij}, x_i)\,, \tag{4.1}$$

donde t_{ij} es el mes j del episodio i, δ_{ij} es el indicador de evento, w_{ij} es el peso asociado al flujo de caja de ese mes y x_i representa el vector de características del conjunto de datos de originación correspondientes al préstamo i.

4.1.1 Construcción de los pesos y tratamiento de casos especiales

La suma de los pesos de cada episodio debe ser igual a 1. Esto es necesario porque, en la adaptación del análisis de supervivencia al problema de la LGD, la función de supervivencia $S(t \mid x)$ se interpreta como la proporción del saldo que todavía queda pendiente de recuperar. Si la suma de pesos no fuera 1, esta interpretación dejaría de ser válida y la LGD estimada no representaría correctamente la fracción del saldo no recuperado. Así, el total de pesos reparte toda la exposición inicial entre lo que se recupera a lo largo del tiempo y lo que queda sin recuperar. Para asegurar que esto se cumple en todos los casos, se aplican tres tratamientos específicos.

En primer lugar, los meses con flujo de caja negativo, presentes en el 0,16 % de las observaciones, se tratan como meses sin recuperación a efectos de modelización. En estos casos se fija `event = 0` y `weight = 0`, ya que dichos importes no representan recuperaciones sino gastos netos o ajustes negativos.

En segundo lugar, los episodios cuya suma de pesos supera la unidad son aquellos en los que las recuperaciones totales exceden el saldo expuesto en el momento del default, lo que daría lugar a una LGD negativa. Este caso afecta a 1.065 episodios. Para poder aplicar el análisis de supervivencia manteniendo la interpretación de los pesos como proporciones de la EAD, los pesos se reescalan proporcionalmente. De esta forma se conserva la distribución temporal de las recuperaciones, pero la recuperación total queda acotada al 100 % de la EAD. Esto equivale a fijar una LGD mínima de cero a efectos de modelización.

En tercer lugar, los episodios cuya suma de pesos es inferior a la unidad reciben una observación artificial en el último mes del episodio. Esta fila tiene `event = 0` y un peso igual a la fracción no recuperada:

$$w_{i,\text{cens}} = 1 - \sum_{j=1}^{m_i} w_{ij}, \tag{4.2}$$

donde m_i es el número de meses observados del episodio i. Esta observación representa la parte del saldo expuesto que no se ha recuperado durante la ventana de observación. El tratamiento se aplica a 20.872 episodios con pérdida parcial o abiertos.

Tras este proceso, la suma de pesos por episodio es exactamente igual a 1 para los 28.502 episodios de la muestra. El dataset de supervivencia resultante contiene 462.840 observaciones, de las cuales 78.856, equivalentes al 17,0 %, corresponden a meses con evento positivo. El resto corresponde a meses sin recuperación o a filas artificiales que recogen la fracción no recuperada.

4.2 Partición Train/Test con Casos Censurados

La división entre conjunto de entrenamiento y conjunto de test se realiza a nivel de episodio, evitando que las observaciones mensuales de un mismo episodio aparezcan simultáneamente en ambos conjuntos. Como se anticipó en el Capítulo 3, los episodios abiertos no pueden utilizarse para validar el error sobre LGD, ya que su pérdida final no se observa. Por ello, la partición train/test se aplica exclusivamente sobre los 27.023 episodios resueltos: el 70 % se asigna al entrenamiento y el 30 % restante al test. Los 1.479 episodios abiertos se incorporan íntegramente al conjunto de entrenamiento, donde contribuyen como observaciones censuradas.

Cuadro 4.1: Partición train/test del dataset de supervivencia

Conjunto	Episodios resueltos	Episodios abiertos	Total episodios	Total filas
Entrenamiento	18.916	1.479	20.395	329.339
Test	8.107	0	8.107	133.501

Esta estrategia evita dos sesgos. En primer lugar, permite aprovechar la información parcial de los episodios abiertos durante el entrenamiento, ya que dichos episodios informan sobre la supervivencia hasta el momento de censura. En segundo lugar, evita evaluar el modelo frente a valores de LGD no observados, lo que introduciría ruido en la validación y podría penalizar artificialmente a los modelos que tratan correctamente la censura.

4.3 Métricas de Evaluación: Discriminación y Calibración

La evaluación de un modelo de LGD requiere medir dos dimensiones complementarias. La primera es la discriminación, que valora la capacidad del modelo para ordenar correctamente los episodios según su pérdida relativa. La segunda es la calibración, que analiza

si el nivel absoluto de las predicciones se aproxima al de las pérdidas observadas. En este trabajo se utilizan cinco métricas que cubren ambas dimensiones: MAE, RMSE, loss shortfall, test t de calibración e índice de concordancia.

4.3.1 Error Absoluto Medio

El Error Absoluto Medio (MAE, por sus siglas en inglés) mide el error medio de predicción en términos absolutos:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| LGD_i - \widehat{LGD}_i \right|. \tag{4.3}$$

Esta métrica es intuitiva y directamente interpretable. Por ejemplo, un MAE de 0,20 indica que la predicción se desvía, en promedio, 20 puntos porcentuales de la LGD observada. Al no elevar los errores al cuadrado, no penaliza de forma desproporcionada los errores extremos, por lo que resulta relativamente robusta frente a la cola de episodios con LGD muy elevada.

4.3.2 Raíz del Error Cuadrático Medio

La Raíz del Error Cuadrático Medio (RMSE, por sus siglas en inglés) se define como:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(LGD_i - \widehat{LGD}_i \right)^2}. \tag{4.4}$$

El RMSE penaliza los errores grandes de forma cuadrática, por lo que es más sensible que el MAE a predicciones muy alejadas de la LGD real. Ambas métricas se interpretan de forma complementaria: una diferencia elevada entre RMSE y MAE indica la presencia de errores extremos en determinados episodios.

4.3.3 Loss Shortfall

El Loss Shortfall (LS) mide el sesgo agregado de calibración del modelo, es decir, si las predicciones tienden sistemáticamente a subestimar o sobreestimar las pérdidas totales de la cartera:

$$LS = 1 - \frac{\sum_{i=1}^{n} \widehat{LGD}_i}{\sum_{i=1}^{n} LGD_i}. \tag{4.5}$$

Un valor positivo indica que el modelo estima unas pérdidas agregadas menores que las reales, lo que desde una perspectiva regulatoria supondría mantener menos capital del necesario. Un valor negativo indica sobreestimación y, por tanto, un exceso de conservadurismo. Un LS cercano a cero refleja una buena calibración global. Esta métrica es

especialmente relevante en gestión del riesgo de crédito, donde un sesgo sistemático puede afectar directamente a los requerimientos de capital y provisiones.

4.3.4 Test t de calibración

El test t de calibración es una aplicación directa del test t de Student para una muestra. Su objetivo es comprobar si la media de los errores de predicción, $LGD_i - \widehat{LGD}_i$, es significativamente distinta de cero. En este contexto, permite evaluar si el modelo presenta un sesgo sistemático, es decir, si tiende a sobreestimar o subestimar la LGD observada.

Formalmente, este test constituye otro método estándar para evaluar la calibración de modelos de LGD (BCE, 2019). La hipótesis nula establece que la LGD predicha es igual en media a la LGD real. El estadístico de contraste se define como:

$$T = \sqrt{N} \cdot \frac{\frac{1}{N}\sum_{i=1}^{N}\left(LGD_i - \widehat{LGD}_i\right)}{s}, \tag{4.6}$$

donde N es el número de observaciones y s es la desviación típica de los errores de predicción:

$$s = \sqrt{\frac{\sum_{i=1}^{N}\left[\left(LGD_i - \widehat{LGD}_i\right) - \frac{1}{N}\sum_{j=1}^{N}\left(LGD_j - \widehat{LGD}_j\right)\right]^2}{N-1}}. \tag{4.7}$$

Bajo la hipótesis nula, el estadístico sigue asintóticamente una distribución t de Student con $N-1$ grados de libertad. Si el p-valor asociado es inferior al nivel de significancia establecido, se rechaza la hipótesis nula y se concluye que el modelo presenta un sesgo de calibración estadísticamente significativo. A diferencia del loss shortfall, que mide el sesgo relativo sobre la cartera agregada, el test t evalúa si dicho sesgo es estadísticamente distinguible del ruido de predicción individual.

4.3.5 Índice de Concordancia

El índice de concordancia, o C-index, mide la capacidad discriminante del modelo. En este trabajo se interpreta como la probabilidad de que, dado un par comparable de episodios, el modelo asigne una LGD predicha mayor al episodio con mayor LGD observada:

$$C = P\left(\widehat{LGD}_i > \widehat{LGD}_j \mid LGD_i > LGD_j\right). \tag{4.8}$$

El índice toma valores entre 0,5 y 1. Un valor de 0,5 equivale a una ordenación aleatoria, mientras que un valor de 1 indica discriminación perfecta. A diferencia del R^2, el C-index no evalúa la precisión en niveles, sino únicamente la ordenación relativa de los episodios. Por ello, resulta útil para analizar si el modelo distingue correctamente entre operaciones de mayor y menor riesgo.

4.4 Optimización de Hiperparámetros

La optimización de hiperparámetros constituye una etapa crítica en el desarrollo de modelos de *machine learning*, ya que estos no se aprenden durante el entrenamiento sino que deben fijarse previamente y condicionan de forma significativa la capacidad predictiva y de generalización del modelo. La optimización de hiperparámetros se realiza exclusivamente sobre el conjunto de entrenamiento, evitando utilizar información del conjunto de prueba durante el proceso de selección, con el objetivo de mejorar su capacidad de predicción en nuevos datos.

La selección de hiperparámetros se realiza mediante validación cruzada sobre el conjunto de entrenamiento, comparando distintas configuraciones a partir de su rendimiento medio en varias particiones internas. En concreto, se ha utilizado validación cruzada con $k = 5$ folds y se restringe el proceso a los episodios resueltos, ya que los episodios abiertos no disponen de una LGD final observable y por tanto, no pueden emplearse para evaluar la métrica de selección. De este modo, el conjunto test queda reservado exclusivamente para la evaluación final del modelo.

En el modelo Cox PH, el ajuste se centró en el penalizador de regularización, que reduce el tamaño de los coeficientes para evitar estimaciones demasiado inestables ante variables correlacionadas. Evaluamos varios niveles de penalización ($\lambda \in \{0{,}001, 0{,}01, 0{,}05, 0{,}1, 0{,}5, 1, 1{,}5, 2\}$), y siguiendo la literatura de los modelos de Cox PH se selecciona aquel valor que maximiza el índice de concordancia [11]. El valor de λ seleccionado se comenta en el Anexo B.

En el caso del *Random Survival Forest*, los hiperparámetros principales incluyen el número de árboles, el número mínimo de observaciones por nodo terminal, el número de variables candidatas evaluadas en cada división y la profundidad máxima de los árboles. Estos parámetros controlan el equilibrio entre sesgo y varianza: árboles más profundos capturan patrones más complejos, pero pueden aumentar el riesgo de sobreajuste; restricciones más fuertes generan modelos más estables, aunque potencialmente menos flexibles. En el Anexo B se muestran los resultados obtenidos para las configuraciones evaluadas y los hiperparámetros seleccionados por maximizar el índice de concordancia para entrenar el modelo de RSF.

4.5 Horizonte de predicción: de la curva de supervivencia a la LGD

La selección del horizonte temporal t^* en el que evaluar la función de supervivencia constituye una de las decisines metodológicas más relevantes de este trabajo. En el análisis de supervivencia clásico, la curva $S(t)$ es en sí misma el resultado: el objetivo es estimar la probabilidad de supervivencia en cada instante. En esta aplicación, en cambio, $S(t \mid X)$ no es el resultado que se busca si no el intrumento para legar a el, evaluando la curva en un momento concreto del tiempo. Es deir, $S(t)$ se utiliza como instrumento para obtener una predicción escalar de la LGD, es decir, un valor entre 0 y 1 que resume la pérdida esperada del episodio.

Para convertir la curva de supervivencia en una predicción de LGD es necesario fijar un horizonte t^* y evaluar la supervivencia en ese punto:

$$\widehat{LGD}_i = S(t^* \mid X_i). \tag{4.9}$$

Esta no es una elección sencilla: un horizonte demasiado corto puede subestimar recuperaciones futuras, mientras que uno excesivamente largo puede evaluar la curva en una zona donde las probabilidades de supervivencia apenas varían, reduciendo así la capacidad discriminante del modelo. Además, la duración del proceso de recobro varía de forma importante entre préstamos, lo que hace cuestionable la elección de un único horizonte global para toda la cartera. Por este motivo, en este trabajo se exploran distintas estrategias de selección del horizonte. Las más relevantes son las siguientes:

- **Horizonte global óptimo.** Consiste en evaluar el modelo para distintos valores de t sobre el conjunto de entrenamiento y seleccionar el horizonte que optimiza un criterio predefinido. Se consideran dos criterios alternativos: minimizar el MAE, que prioriza la precisión individual préstamo a préstamo, y minimizar $|LS|$, que prioriza la calibración agregada de la cartera.Como horizonte base para la comparación entre el modelo de Cox y el *Random Survival Forest*, se seleccionara un t* que combina ambas dimensiones: se minimiza —LS— dentro del conjunto de valores de t para los que el MAE se mantiene próximo a su mínimo, garantizando así un compromiso entre la calibración agregada de la cartera y la precisión individual de las predicciones.

- **Horizonte individualizado por percentil de $S(t \mid X)$.** Dado que cada préstamo puede presentar un ritmo de recuperación distinto, se define un horizonte específico para cada operación:

  $$t_i^* = \text{mín}\left\{t : S(t \mid X_i) \leq 1 - p\right\}. \tag{4.10}$$

  Este valor representa el primer mes en que la curva de supervivencia del préstamo ha caído un porcentaje p respecto a su nivel inicial. El parámetro p se selecciona mediante búsqueda en rejilla sobre el conjunto de entrenamiento, utilizando el MAE como criterio de optimización.

- **Horizonte óptimo por decil de LGD predicha.** Este enfoque refina el anterior permitiendo que el porcentaje p varíe según el perfil de riesgo del préstamo. Para ello, se agrupan las observaciones por deciles de LGD predicha y se selecciona, en cada decil, el valor de p que minimiza el MAE. La idea es que préstamos con mayor LGD esperada, normalmente asociados a procesos de recuperación más lentos, pueden requerir horizontes más largos.

- **Horizonte aprendido mediante regresión.** Como alternativa, se entrena un modelo de regresión para predecir directamente el tiempo de recuperación relevante de cada préstamo. Este tiempo se define como el primer mes en que se ha acumulado el un porcentaje concreto de los flujos de caja observados.

- **Enfoque híbrido.** Por último, se propone combinar el horizonte individualizado por decil con el horizonte global. Para los préstamos en los que el enfoque por decil ofrece mejores resultados se utiliza el horizonte específico de su grupo, mientras que

para el resto se mantiene el horizonte global t^*. La frontera entre ambos grupos se define a partir de la LGD predicha por el modelo, y el umbral que separa las dos estrategias se determina empíricamente sobre el conjunto de entrenamiento.

4.6 Calibración Post-hoc de Modelos

La calibración *post-hoc* consiste en ajustar las predicciones de un modelo ya entrenado para corregir sesgos sistemáticos, sin modificar sus parámetros internos. En modelos de riesgo de crédito es especialmente relevante porque una buena ordenación de los préstamos no garantiza que el nivel agregado de pérdida esté bien calibrado (BCE, 2019).

En este trabajo se aplica una calibración aditiva por tramos, similar a una versión discreta de la calibración isotónica. Para ello, se divide la distribución de LGD predicha en el conjunto de entrenamiento en quintiles y, dentro de cada tramo k, se calcula el sesgo medio $\bar{\varepsilon}_k$ entre la LGD observada y la predicha. Esta corrección se aplica después a las predicciones del conjunto de test, acotando el resultado al intervalo $[0, 1]$:

$$\widehat{LGD}_i^{cal} = \text{clip}\left(\widehat{LGD}_i + \bar{\varepsilon}_k,\ 0,\ 1\right). \tag{4.11}$$

La corrección se estima únicamente con información del conjunto de entrenamiento, por lo que el test permanece reservado para la evaluación . Finalmente, las predicciones calibradas se coparan con las originales mediante las métricas de la Sección 4.3, con el objetivo de comprobar si mejora la precisión económica sin deteriorar la capacidad discriminante del modelo.

Capítulo 5

Resultados

5.1 Análisis descriptivo: estimador de Kaplan-Meier

El estimador de Kaplan-Meier proporciona una primera aproximación no paramétrica al proceso de recuperación. Antes de presentar los resultados, conviene recordar la conexión entre la función de supervivencia y la LGD establecida en la Sección 2.2.4. En este trabajo, el evento de interés no es único ni binario como en las aplicaciones clásicas del análisis de supervivencia, sino que cada mes en que se produce un *cash flow* positivo constituye un evento parcial ponderado por su peso económico relativo. Así, $S(t)$ representa la proporción del saldo expuesto que el modelo estima que no habrá sido recuperada hasta el mes t, de forma que la LGD se aproxima directamente como:

$$\widehat{LGD}_i = S(t^* \mid X_i) \tag{5.1}$$

El estimador de Kaplan-Meier ofrece una primera estimación de esta función sin incluir covariables y constituye, por tanto, la referencia descriptiva sobre la que se construirán los modelos predictivos de los apartados siguientes.

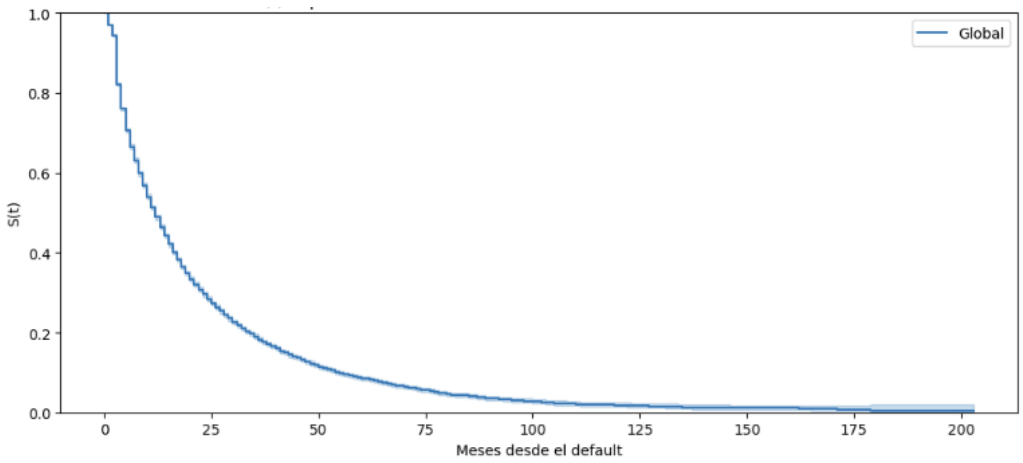

Figura 5.1: Curva global de supervivencia de Kaplan-Meier

La curva global, representada en la Figura 5.1, muestra una caída pronunciada en los primeros meses tras el default: aproximadamente el 25 % de los episodios genera su primer flujo de caja antes del mes 5, y la curva se estabiliza por encima de cero a partir del mes 75. Esto refleja la existencia de una fracción de episodios que no llega a generar recuperación alguna durante el periodo de observación. Esta forma característica, con una caída inicial rápida seguida de una cola larga, es coherente con la distribución asimétrica de la duración de los episodios descrita en la Sección 3.4.3.

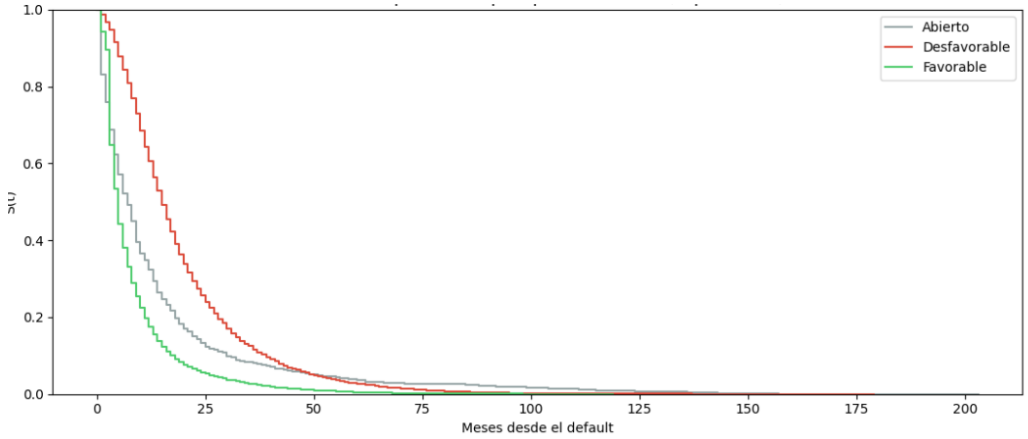

Figura 5.2: Curvas de supervivencia de Kaplan-Meier por tipo de resolución

El análisis por tipo de terminación, recogido en la Figura 5.2, revela diferencias estructurales entre grupos que la curva global oculta. Los episodios con terminación favorable presentan una caída muy pronunciada en los primeros meses —la curva cae por debajo de 0,20 antes del mes 10 y se aproxima a cero hacia el mes 30—, lo que indica que la mayor parte de las recuperaciones en este grupo se concentra en una ventana temporal muy corta. Los episodios desfavorables en cambio, mantienen valores de S(t) superiores a 0,40 hasta el mes 25 y no alcanzan valores próximos a cero hasta pasado el mes 75, lo que refleja la duración extendida de los procesos con terminación desfavorable. Los episodios abiertos presentan un comportamiento intermedio, resultado de mezclar episodios con distintos perfiles de riesgo y una ventana de observación más corta por construcción.

La diferencia entre las curvas de episodios favorables y desfavorables es estadísticamente significativa según el test log-rank (estadístico = 91.743,7; $p < 0,001$), confirmando que ambos grupos siguen procesos de recuperación sustancialmente distintos. Este resultado confirma que los distintos grupos presentan distintas dinámicas de recuperación. Aunque el tipo de terminación no es observable en el momento del default, los modelos predictivos desarrollados en los apartados siguientes intentarán capturar indirectamente esta heterogeneidad a partir de las características de originación del préstamo.

5.2 Modelo de regresión base

El modelo OLS se estima como *benchmark* previo a los modelos de análisis de supervivencia, utilizando las variables seleccionadas en el Capítulo 3 e incorporando términos de interacción. Cabe señalar que dado que el desarrollo de este modelo no constituye el objetivo central del trabajo sino una referencia de comparación, se utiliza una versión sencilla. Enfoques de regresión más elaborados podrían mejorar los resultados aquí obtenidos, pero su desarrollo queda fuera del alcance de este estudio. Se evaluó también una versión calibrada *post-hoc* mediante corrección aditiva por tramos, que mejoró marginalmente las métricas agregadas, pero se optó por reportar el modelo sin calibrar para facilitar la comparación directa con los modelos de supervivencia en la Sección 5.4. Los coeficientes completos del modelo se recogen en el Anexo B.

En general, los coeficientes presentan los signos esperados, y el tipo de interés se mantiene como la variable de mayor impacto individual, con un coeficiente de 0,556. La interacción entre el tipo de interés vigente en el momento del default y la antigüedad del préstamo, $rate \times loan\_age$, resulta ser la más significativa entre todas las iteraciones que se estudiaron (coeficiente $= -0,276$; $p < 0,001$). Este resultado indica que el efecto positivo del tipo de interés sobre la LGD crece conforme el préstamo envejece.

La evaluación sobre el conjunto de test muestra un MAE de 0,206, un *Loss Shortfall* de $-0,014$ y un p-valor de 0,219 en el test t de calibración. Estos resultados sugieren que el modelo está razonablemente calibrado en media y que no presenta un sesgo sistemático estadísticamente significativo. La Figura 5.3 ilustra la principal limitación del modelo: la distribución predicha es unimodal y centrada en torno a 0,30, sin reproducir la marcada bimodalidad de la LGD real. El modelo sobreestima en los tramos bajos y subestima en los altos, pero ambos sesgos se compensan al agregar la cartera — lo que explica la aparente contradicción entre un R^2 modesto y un LS casi nulo. Esta propiedad es una consecuencia directa de la regresión a la media: al minimizar el error cuadrático, el modelo es atraído hacia la media de la distribución independientemente del perfil real de cada préstamo, garantizando una buena calibración agregada pero no una buena discriminación individual. Esta limitación, inherente a la forma funcional lineal, motiva la exploración del uso de los modelos de análisis de supervivencia en los apartados siguientes.

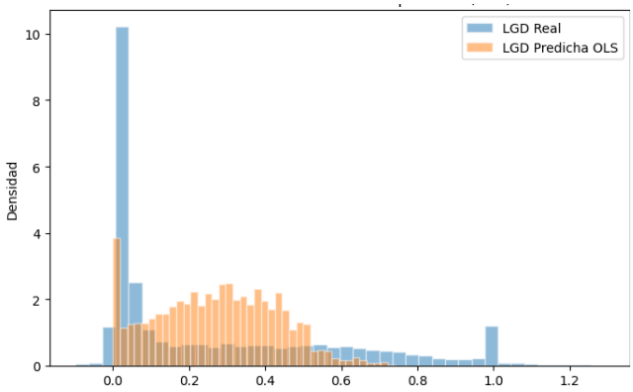

Figura 5.3: Distribución de la LGD observada y predicha por el modelo OLS

5.3 Modelos de Análisis de Supervivencia

En esta sección se presentan los resultados de los dos modelos de análisis de supervivencia desarrollados en este trabajo. Ambos estiman una función de supervivencia individual, $S(t \mid X_i)$, para cada préstamo y derivan la LGD evaluándola en un horizonte temporal t^* seleccionado empíricamente, como semencionó en la Sección 4.5. A diferencia del OLS, ambos modelos incorporan los 1.479 episodios abiertos durante el entrenamiento como observaciones censuradas, aprovechando toda la información disponible. Este apartado se organiza en tres bloques: en primer lugar se presentan los resultados del Cox PH, en segundo lugar los del RSF y, finalmente, se analiza de forma conjunta el efecto de la elección del horizonte temporal sobre el rendimiento de ambos modelos.

5.3.1 Modelo Cox PH

La Figura 5.4 muestra la curva de supervivencia estimada por el modelo de Cox para un episodio concreto. Esta curva, $S(t \mid X_i)$, representa la fracción del saldo que el modelo estima que no habrá sido recuperada hasta cada mes t: parte de 1 en el momento del default y decrece conforme el modelo espera que se vayan produciendo recuperaciones. Para obtener una predicción de LGD a partir de esta curva es necesario evaluarla en un punto concreto, que en este modelo se ha fijado en $t^* = 27$ meses por ser el valor que ofrece el mejor equilibrio entre minimizar el MAE y el LS sobre el conjunto de entrenamiento, tal y como se describe en la Sección 4.5. Este punto se marca con la línea vertical de puntos en la figura: el valor de la curva en ese instante, $S(27 \mid X_i) = 0,267$, es la LGD predicha para este préstamo, frente a una LGD real de 0,173. Se aprecia además que la curva sigue cayendo más allá de $t^* = 27$: si se hubiera elegido un horizonte más tardío, la predicción habría sido menor, lo que ilustra el riesgo de subestimar la LGD cuando el horizonte se fija demasiado lejos.

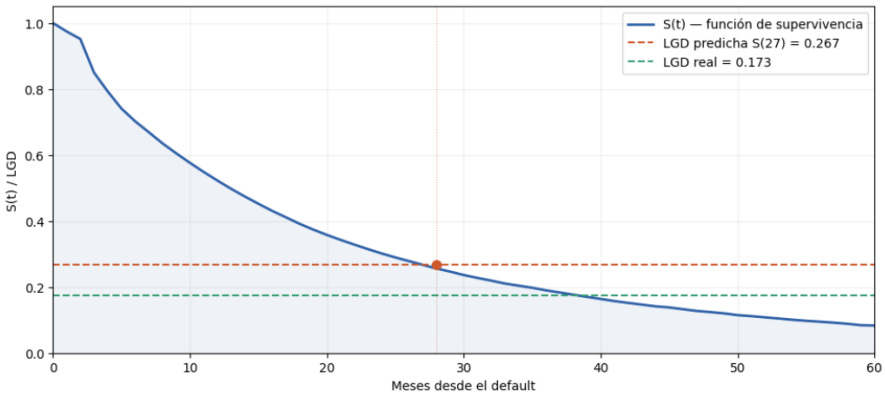

Figura 5.4: Predicción de LGD a partir de la curva de supervivencia del modelo Cox PH

El test de Schoenfeld revela que la mayoría de las variables del modelo violan el supuesto de riesgos proporcionales ($p < 0,05$), lo que implica que el efecto de las covariables sobre el *hazard rate* no es constante a lo largo del tiempo. Como se anticipó en la Sección 2.2.3, esta situación es habitual en aplicaciones reales del modelo de Cox y no invalida el modelo

como herramienta predictiva. Los coeficientes siguen siendo informativos en cuanto a la dirección del efecto, aunque su interpretación en términos de magnitud debe hacerse con cautela . Esta limitación es una de las motivaciones principales para el uso del RSF en el apartado siguiente, que no impone ningún supuesto sobre la forma del *hazard rate*.

Por otro lado, los coeficientes del modelo recogen el efecto de cada variable sobre el *hazard rate* de recuperación, es decir, sobre la velocidad instantánea con que el préstamo genera flujos de caja. Por tanto, su interpretación es inversa a la del OLS: un coeficiente negativo implica menor velocidad de recuperación y, en consecuencia, mayor LGD esperada. Aunque los resultados completos se recogen en el Anexo B, cabe mencionar que la variable de mayor impacto es `current_interest_rate_ep_ord` (coef. $= -1,225$): los préstamos con tipos de interés más altos presentan una velocidad de recuperación un 70,6 % inferior a la de los préstamos con tipos bajos. Le siguen `original_loan_term_ord` y `ltv_ord` ambas con el signo negativo esperado. Por su parte, `loan_age_ord` presenta un coeficiente positivo (coef. $= 0,370$), indicando que los préstamos más antiguos en el momento del default se recuperan más rápidamente, efecto que el OLS capturaba de forma indirecta a través de la interacción con el tipo de interés.

Los resultados del modelo según la elección del $t^* = 27$ mencionada anteriormente muestran que el Cox PH está muy bien calibrado a nivel de cartera: un *Loss Shortfall* de $-0,001$ y un p-valor de 0,910 en el test t indican que la suma de las LGD predichas coincide con la suma de las LGD reales, sin sesgo sistemático significativo. Sin embargo, su capacidad para distinguir entre préstamos de alto y bajo riesgo es más limitada — un índice de concordancia de 0,647 implica que el modelo ordena correctamente un par de préstamos según su LGD relativa en algo más de seis de cada diez casos —, y el error individual, con un MAE de 0,213, sigue siendo considerable en una escala acotada entre 0 y 1. El modelo resulta por tanto adecuado para fines de calibración de cartera, pero su capacidad discriminante a nivel de préstamo individual deja margen de mejora, que se explora con el RSF en el apartado siguiente.

5.4 *Random Survival Forest*

El RSF se entrena sobre el mismo conjunto que el Cox PH, con las 28 covariables descritas en la Sección 3.5, sin necesidad de discretización ni de normalización *min-max* —ventaja de los modelos basados en árboles. La configuración final, seleccionada mediante validación cruzada (Sección 4.4), corresponde a 200 árboles, un mínimo de 20 observaciones por hoja terminal y un 50 % de variables candidatas evaluadas en cada división.

Al igual que en el Cox PH, la LGD se obtiene evaluando la función de supervivencia estimada en un horizonte t^*, y de la misma forma t^* se selecciona como el valor que minimiza $|LS|$ dentro del conjunto de valores de t para los que el MAE se mantiene próximo a su mínimo sobre el conjunto de entrenamiento, resultando en $t^* = 27$ meses (el mismo horizonte seleccionado para el Cox PH). Cabe señalar que esta coincidencia es un resultado empírico de este conjunto de datos y no una propiedad general: ambos modelos seleccionan t^* de forma independiente sobre el conjunto de entrenamiento, pudiéndose obtener horizontes distintos.

La Figura 5.5 ilustra el proceso de predicción para un préstamo concreto. Cada uno de los 200 árboles del bosque genera su propia curva de supervivencia a partir de las observaciones que caen en su nodo terminal correspondiente —en la figura se muestran en verde claro una muestra de 30 de estas curvas individuales—, y el RSF promedia las 200 curvas para obtener la curva final $S(t \mid X_i)$, representada en azul oscuro. La LGD predicha se obtiene leyendo esta curva promedio en $t^* = 27$: en este ejemplo, $S(27 \mid X_i) = 0,325$, frente a una LGD real de 0,300. A diferencia del Cox PH, donde la forma de la curva está determinada por la función paramétrica $S_0(t)^{\exp(x'\beta)}$, aquí la curva surge directamente de los datos observados en los nodos del árbol, lo que le permite adaptarse a formas más irregulares —como se aprecia en el tramo final, donde $S(t)$ se estabiliza de forma escalonada en lugar de decrecer suavemente.

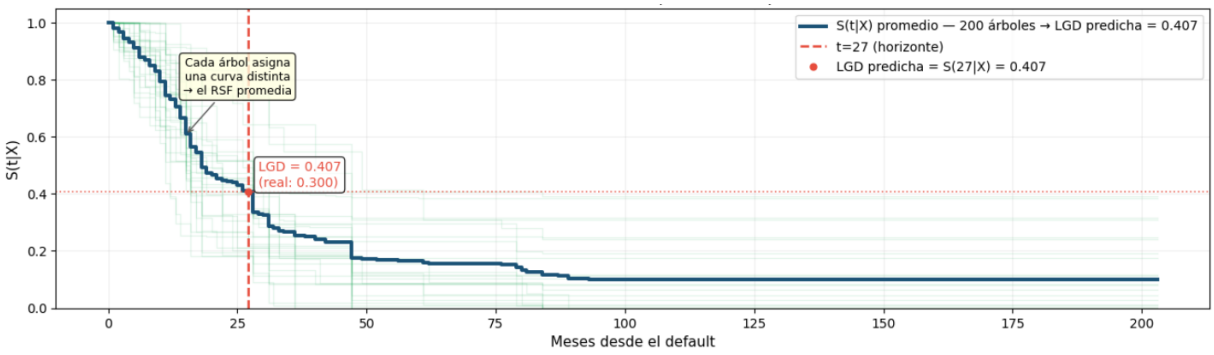

Figura 5.5: Predicción del modelo RSF para un préstamo concreto

Para situar este ejemplo en el contexto del conjunto de test, la Figura 5.6 compara la curva de supervivencia de este mismo préstamo con la del conjunto de test en su totalidad. La línea verde representa la curva media de todos los préstamos del test, y la banda sombreada recoge el intervalo entre los percentiles 25 y 75 de las curvas individuales en cada instante t —es decir, el rango donde se sitúa la mitad central de los préstamos del test—. La curva del préstamo seleccionado (en rojo) se mantiene por encima de la media y de la banda sombreada durante los primeros 50 meses, lo que indica que, según el modelo de RSF, este préstamo recupera algo más lentamente que la mayoría de la cartera en ese tramo inicial.

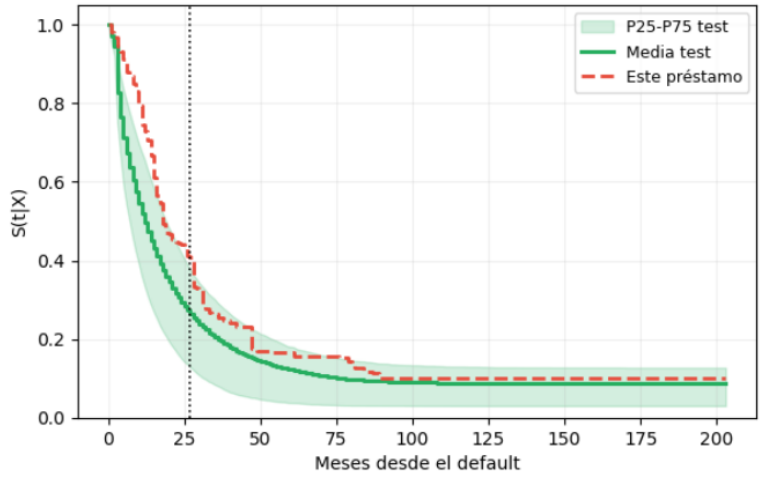

Figura 5.6: Curva de supervivencia del préstamo seleccionado frente a la distribución del conjunto de test en el modelo RSF

En cuanto a la importancia de las variables del modelo, el VIMP (Sección 2.4) confirma a `current_interest_rate_ep_ord` (VIMP = 0,0392) como la variable más relevante, muy por encima del resto, lo que resulta coherente con su importante papel en el Cox PH y en el análisis univariante de la Sección 3.5. Le sigue `loan_age_ord` (VIMP = 0,0083), confirmando que la antigüedad del préstamo en el momento del default es, junto con el tipo de interés, el factor más relevante para predecir la LGD. El resto de variables presentan importancias considerablemente menores (VIMP $< 0,004$). De las 30 variables del modelo, 19 presentan importancia positiva y 11 negativa. El listado completo se recoge en el Anexo B.

La evaluación sobre el conjunto de test con $t^* = 27$ presenta un MAE de 0,192 y un RMSE de 0,237, ambos inferiores a los del Cox PH (0,213 y 0,271), lo que indica una mejora en la precisión individual de las predicciones. El *Loss Shortfall* es de 0,014, con un p-valor de 0,219 en el test t, por lo que el modelo está bien calibrado a nivel de cartera, de forma similar al Cox PH. El índice de concordancia es de 0,684, superior al 0,647 del Cox PH, lo que confirma una mejora sustancial en la capacidad discriminante. En conjunto, bajo un horizonte común de $t^* = 27$, el RSF mejora al Cox PH en las tres dimensiones evaluadas —error individual, calibración y discriminación—, coherente con su mayor flexibilidad para capturar relaciones no lineales e interacciones entre variables sin necesidad de especificarlas a priori.

5.5 Comparativa final de modelos estáticos

La Figura 5.7 muestra la distribución de LGD predicha por el RSF en test frente a la LGD real. A diferencia del histograma del OLS, cuya distribución predicha era claramente unimodal, el RSF sí logra capturar parte de la bimodalidad real: aparece una concentración de predicciones bajas, coherente con el grupo de episodios favorables, y una cola que llega hasta valores altos. Sin embargo, es cierto también que la mejora es solo parcial, pues la mayor parte de las predicciones sigue concentrada en la zona central.

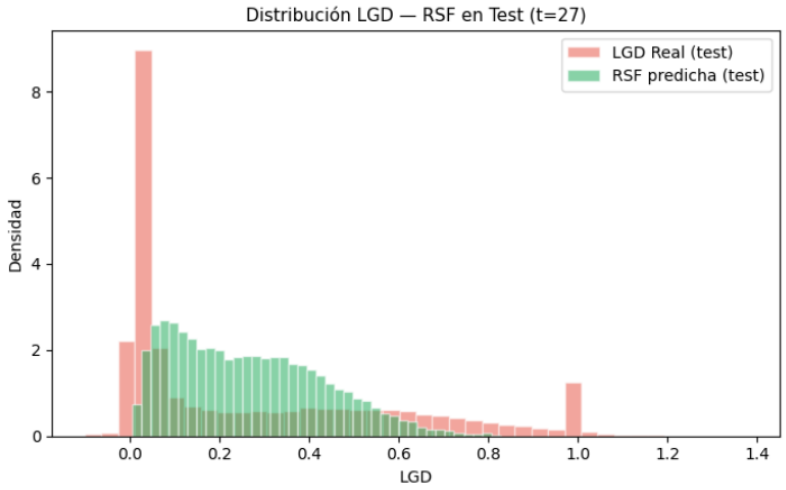

Figura 5.7: Distribución de la LGD observada y predicha por el modelo RSF en el conjunto de test

En conjunto, los tres modelos muestran una jerarquía clara: el RSF mejora al Cox PH y al OLS tanto en error individual como en capacidad de discriminación, manteniendo una calibración de cartera similar. Esto indica que el uso de *Machine Learning* aporta algo real frente a los enfoques clásicos, tanto frente a la regresión lineal como frente al Cox PH, pero con el mismo inconveniente de siempre: mayor capacidad predictiva a cambio de menor interpretabilidad. Conviene aclarar, sin embargo, que este trabajo no pretende demostrar que el análisis de supervivencia sea mejor que la regresión, ni que el RSF sea la mejor técnica posible para este problema. Tanto el modelo de regresión como los modelos de supervivencia podrían mejorarse bastante — con regresión beta, calibraciones más elaboradas o aprovechando mejor los episodios censurados —, y esa comparación queda fuera del alcance de este trabajo. El objetivo aquí es más simple: mostrar que el análisis de supervivencia es un enfoque aplicable y con potencial para modelizar la LGD, capaz de aprovechar información que los enfoques tradicionales simplemente descartan.

Por último, ninguno de los tres modelos reproduce bien la bimodalidad de la LGD real, lo que sugiere que parte del margen de mejora no depende solo del modelo elegido sino también de otras decisiones, en particular de la elección del horizonte temporal t^*. Hasta ahora se ha usado $t^* = 27$ como horizonte fijo, por ser el valor que minimiza $|LS|$ dentro del rango donde el MAE es mínimo en train, pero no existe un criterio único aceptado en la literatura para esta elección. A continuación se analizan los resultados de las alternativas presentadas en la Sección 4.5, para ver si elegir t^* de otra forma podría mejorar el rendimiento del RSF.

5.6 Selección del horizonte temporal

Sobre el RSF se probaron las distintas estrategias de selección de t^* descritas en la Sección 4.5. Además de los enfoques que se comparan a continuación, se probó también un horizonte aprendido mediante un modelo de regresión que predecía directamente la duración esperada de cada préstamo, siendo los resultados especialmente poco satisfactorios

—horizontes muy alejados de los valores reales y, con ello, errores extremos en varios tramos—, por lo que se descartó y no se incluye en la comparación.

La Figura 5.8 muestra el MAE por tramo de LGD real para los tres enfoques restantes —horizonte por percentil individual, horizonte por decil de LGD predicha e híbrido— junto con el horizonte fijo $t^* = 27$ como referencia. El horizonte por percentil individual asigna a cada préstamo el momento en que su curva $S(t \mid X_i)$ se estabiliza, y resulta ser muy optimista: predice LGD bajas de forma generalizada, lo que explica su buen resultado en el tramo 0–10 % (MAE = 0,045) pero su pésimo desempeño en el resto, llegando a duplicar el error del horizonte fijo en el tramo $> 100\,\%$ (0,893 frente a 0,639) y con un sesgo de cartera inaceptable (LS = 0,66). El horizonte por decil sigue un patrón similar, aunque más moderado. El híbrido —que combina el horizonte por decil en los préstamos de menor riesgo con el horizonte fijo en el resto— ofrece el mejor equilibrio entre tramos, aunque tampoco logra superar al horizonte fijo en los tramos de LGD alta.

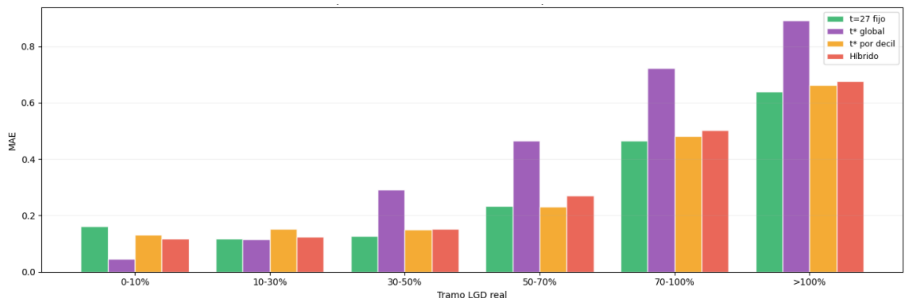

Figura 5.8: MAE por tramo de LGD real según la estrategia de selección del horizonte temporal

En conjunto, ningún enfoque mejora al horizonte fijo en todos los tramos a la vez: las mejoras en un extremo de la distribución se consiguen siempre a costa de empeorar en el otro. Aunque estos resultados se presentan solo para el RSF, el mismo problema aparece en el Cox PH, ya que ambos modelos comparten la misma limitación de fondo: la curva $S(t \mid X)$ describe igual de bien (o de mal) a préstamos con ritmos de recuperación muy distintos, y ningún horizonte —fijo o individualizado mediante las estrategias exploradas— corrige esto de forma satisfactoria.

Esto apunta a que la mejora más prometedora no está en cómo se lee la curva $S(t \mid X)$, sino en predecir directamente cuánto va a tardar cada préstamo en resolverse. Como muestra de ello, la Figura 5.9 recoge el resultado de una prueba exploratoria en la que se incluyó `duration_months` —la duración real del episodio— como covariable adicional del RSF, lo que mejora notablemente las métricas de ajuste. Este resultado no puede usarse directamente, ya que `duration_months` no se conoce en el momento del default y su uso constituiría *data leakage*; sin embargo, sugiere que si se lograra estimar de forma fiable la duración esperada del proceso de recobro a partir de variables disponibles en origen, esa información podría mejorar sustancialmente la predicción de la LGD. Esta idea se retoma como propuesta de trabajo futuro.

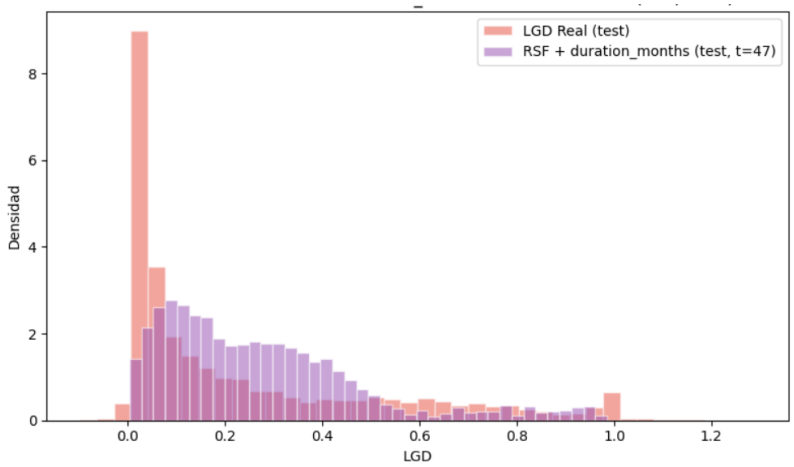

Figura 5.9: Distribución de la LGD observada y predicha por el RSF incluyendo la duración real del episodio

5.7 Predicción dinámica de la LGD residual

Hasta ahora la LGD se ha estimado en el momento del default, asignando a cada préstamo una predicción fija que no se actualiza posteriormente. Esta sección presenta los resultados de la extensión propuesta en la Sección 2.2.4: estimar, para préstamos que llevan m meses en mora, la pérdida residual esperada a partir de ese momento mediante:

$$LGD_{residual}(m) = \frac{S(m + t^* \mid X_i)}{S(m \mid X_i)}. \tag{5.2}$$

Al igual que en el apartado anterior, y con el objetivo de mantener el foco en la idea principal de esta extensión, los resultados se presentan únicamente para el RSF. La formulación de la propuesta es idéntica para el Cox PH, al igual que la interpretación de los resultados, ya que ambos modelos presentan un comportamiento similar.

La Figura 5.10 muestra la evolución del MAE y el RMSE de la LGD residual predicha por el RSF para cada cohorte de préstamos que llevan m meses en mora. Ambas métricas presentan un pico pronunciado en $m = 3$ (RMSE = 0,388), que coincide con el punto de mayor desajuste entre la LGD residual real y la predicha. A partir de $m = 6$, sin embargo, ambas métricas se estabilizan en torno a 0,26 (MAE) y 0,33 (RMSE) durante el resto del *workout period*. Que el error se mantenga prácticamente constante a partir de ese punto es un resultado positivo: sugiere que, una vez superados los primeros meses de mora, la capacidad predictiva del modelo no se degrada de forma significativa con el tiempo transcurrido en default.

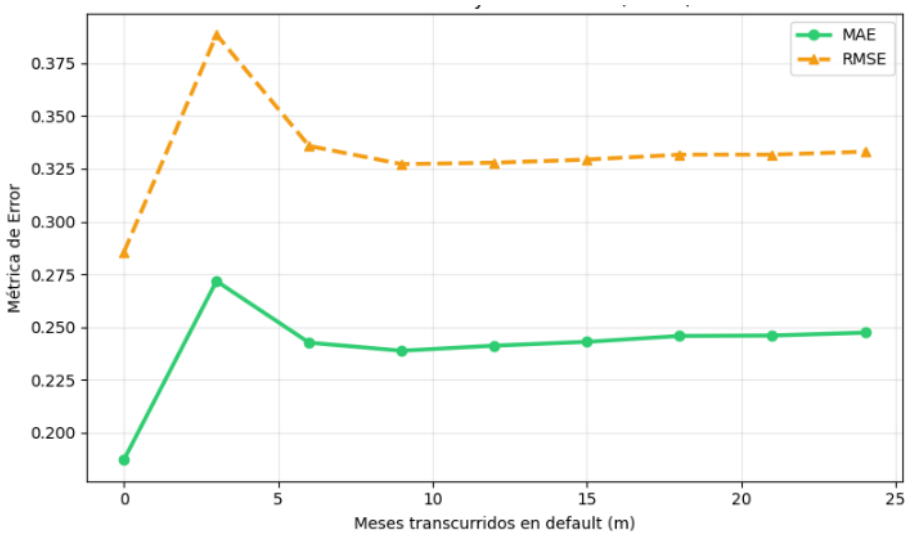

Figura 5.10: Evolución del error de predicción de la LGD residual por meses en mora

En conjunto, estos resultados muestran que la predicción dinámica de la LGD residual es una extensión viable del enfoque de supervivencia, aunque con un nivel de error similar al de la predicción en el momento del default. La mejora del modelo en este contexto pasaría por las mismas ideas discutidas en la Sección 5.3.4 — especialmente, una mejor caracterización del tiempo de recuperación de cada préstamo —, lo que confirma que esta propuesta, aunque prometedora, requeriría un desarrollo más profundo para alcanzar su potencial.

Capítulo 6

Conclusiones y trabajo futuro

6.1 Conclusiones

Este trabajo no nace con la idea de presentar un modelo de LGD listo para sustituir a los que ya se usan en la práctica bancaria, sino de plantear una pregunta más sencilla: ¿qué pasa si dejamos de tratar la LGD como un número fijo que solo conocemos al final, y la entendemos como un proceso que va ocurriendo en el tiempo? El análisis de supervivencia ofrece justo ese punto de vista. En este contexto, descartar los episodios abiertos supone perder información precisamente sobre los casos más difíciles de modelizar. El análisis de supervivencia ofrece una alternativa natural, ya que permite incorporar estos episodios como observaciones censuradas y modelizar explícitamente la dinámica temporal del proceso de recuperación.

La primera conclusión es que el enfoque propuesto resulta viable y útil para este problema. A partir de los flujos de caja observados se ha construido una variable objetivo coherente con la definición económica de LGD, se han transformado los episodios de default en un dataset de supervivencia ponderado por exposición, y se ha estimado una función de supervivencia interpretable como la proporción del saldo aún no recuperada en cada horizonte temporal. Esta formulación permite conectar de forma directa la salida de los modelos de supervivencia con una predicción de LGD, evaluando la curva individual $S(t \mid X_i)$ en un horizonte t^*.

Los resultados empíricos muestran que los modelos de supervivencia son competitivos frente al modelo de regresión lineal utilizado como *benchmark*. El modelo Cox PH alcanza una calibración agregada muy precisa, con un *Loss Shortfall* prácticamente nulo, aunque su capacidad de discriminación individual queda limitada por ser un modelo semiparammetrico y por la violación del supuesto de riesgos proporcionales en la mayoría de covariables. Aun así, sus coeficientes ofrecen unainterpretabilidad económica útil: variables como el tipo de interés, el LTV o el plazo original del préstamo afectan de forma clara a la velocidad esperada de recuperación y, por tanto, a la LGD esperada.

El *Random Survival Forest* mejora al Cox PH y al OLS en error individual y capacidad de ordenación, manteniendo una calibración de cartera similar. Este resultado sugiere que la combinación entre análisis de supervivencia y aprendizaje automático permite

capturar relaciones no lineales e interacciones entre variables que los modelos clásicos no recogen de forma suficiente. La mejora, sin embargo, es moderada y no elimina por completo una de las principales dificultades del problema: la distribución real de la LGD es bimodal, mientras que las predicciones de los modelos tienden a concentrarse en valores intermedios. Esta regresión hacia la media aparece tanto en el OLS como en los modelos de supervivencia, aunque con distinta intensidad.

La selección del horizonte temporal t^* se revela como una decisión central. El horizonte fijo elegido empíricamente ofrece un buen equilibrio entre error individual y calibración agregada, pero los resultados muestran que no existe una elección universalmente óptima: desplazar el horizonte mejra unas zonas de la distribución a costa de empeorar otras. Las estrategias individualizadas exploradas, basadas en percentiles o deciles de la curva de supervivencia, no corrigen de forma satisfactoria este problema. Esto indica que la limitación no está solo en cómo se lee la curva $S(t \mid X_i)$, sino en la dificultad de anticipar correctamente el ritmo de recuperación de cada préstamo.

Por último, la extensión dinámica de la LGD residual muestra que el marco de supervivencia puede utilizarse no solo para estimar la pérdida esperada en el momento del default, sino también para actualizarla cuando el préstamo ya lleva varios meses en mora. Aunque el nivel de error sigue siendo similar al de la predicción estática, su posible aplicación abre la puerta a una estimación más operativa de la LGD, adaptada al seguimiento mensual de carteras en default. De esta manera, el trabajo confirma que el análisis de supervivencia es un enfoque aplicable y con potencial para la modelización de LGD, especialmente cuando existen observaciones censuradas y se desea aprovechar la dimensión temporal del proceso de recuperación.

6.2 Limitaciones

La principal limitación del estudio es que los modelos se han desarrollado sobre una muestra concreta del dataset de Freddie Mac y bajo una definición específica de episodio de default, recuperación y censura. Aunque esta construcción es coherente con los objetivos del trabajo, otras decisiones de segmentación, descuento de flujos o tratamiento de episodios curados podrían modificar parcialmente los resultados.

Una segunda limitación es que la comparación con modelos de regresión se ha planteado como referencia, no como una competición exhaustiva entre familias de modelos. El OLS utilizado es deliberadamente sencillo, por lo que no permite concluir que el análisis de supervivencia sea superior a cualquier enfoque alternativo de regresión. Modelos más complejos, desarrollados con un trabajo específico de ajuste y validación, podrían mejorar la predicción estática de LGD y deberían compararse en igualdad de condiciones en trabajos posteriores.

También debe señalarse que la información disponible en el momento del default en la fuente de datos del trabajo, no parece suficiente para reproducir completamente la heterogeneidad observada en la LGD final. La prueba exploratoria que incorpora la duración real del episodio mejora sustancialmente la forma de la distribución predicha, pero dicha variable no puede utilizarse en un entorno real porque no se conoce ex ante. Este resul-

tado sugiere que una parte importante del error procede de la dificultad de anticipar la duración y el ritmo del proceso de recobro.

Finalmente, conviene reconocer que este trabajo aborda un problema mucho más amplio de lo que 60 páginas permiten desarrollar en profundidad, y varias decisiones tomadas durante la construcción de la muestra y de los modelos podrían revisarse en un estudio posterior. Por ejemplo las estrategias de elección de t^* se han evaluado de forma exploratoria. Aunque ayudan a entender el comportamiento de los modelos, no constituyen un procedimiento cerrado para usarlo en regulación o en la práctica.

6.3 Trabajo futuro

Una línea natural de trabajo futuro sería desarrollar modelos específicos para predecir la duración esperada del proceso de recuperación. Los resultados obtenidos sugieren que mejorar la estimación del tiempo hasta la resolución podría ser más útil que seguir refinando únicamente la elección del horizonte t^*. En este trabajo, dicho horizonte se fija de forma externa y se utiliza como punto de lectura de la curva de supervivencia, pero el análisis muestra que esta decisión condiciona de manera importante la LGD estimada. La propuesta sería, por tanto, modelizar explícitamente el tiempo de resolución esperado de cada episodio y utilizar esa predicción para adaptar el horizonte de evaluación a las características individuales del préstamo.

Esta línea permitiría conectar de forma más directa la dinámica temporal del default con la severidad final de la pérdida. En particular, podrían explorarse modelos de duración orientados a anticipar si un episodio se resolverá de forma temprana o prolongada, así como enfoques conjuntos que combinen la predicción del tiempo de recuperación con la estimación de la LGD residual. De este modo, el horizonte temporal dejaría de ser un parámetro fijo elegido en validación y pasaría a formar parte del propio sistema predictivo, lo que podría mejorar tanto la interpretación económica del modelo como su utilidad práctica para el seguimiento de carteras en default.

Apéndice A

Anexo A: tablas complementarias

Este anexo recoge las tablas complementarias utilizadas en la construcción de la muestra, el cálculo de la LGD y el proceso de selección e ingeniería de variables. Su objetivo es mantener el cuerpo principal del trabajo centrado en la metodología y los resultados principales, trasladando aquí el detalle tabular de apoyo.

A.1 Variables originales utilizadas

Las siguientes variables proceden directamente de los ficheros de originación y *performance* del *Single-Family Loan-Level Dataset* de Freddie Mac. Se documentan únicamente aquellas que intervienen directamente en la construcción de las variables del análisis o en el cálculo de la LGD, con el objetivo de facilitar la trazabilidad del proceso. Las definiciones se basan en la guía oficial del dataset (Freddie Mac, 2026).

A.1.1 Fichero de originación

Cuadro A.1: Variables originales del fichero de originación utilizadas en el análisis

Variable	Tipo	Descripción
`credit_score`	Numérica entera	Puntuación crediticia del prestatario en el momento de la originación, generalmente en el rango 300–850. Valores de 9999 indican no disponible.
`ltv`	Numérica entera	Ratio préstamo-valor original (*Loan-to-Value*): cociente entre el importe del préstamo y el valor de tasación o precio de compra del inmueble en el momento de la originación, expresado en porcentaje.

Continúa en la página siguiente

Variable	Tipo	Descripción
`cltv`	Numérica entera	Ratio préstamo-valor combinado (*Combined LTV*): incluye en el numerador el importe del préstamo más cualquier financiación secundaria comunicada por el vendedor.
`dti`	Numérica entera	Ratio deuda-ingresos (*Debt-to-Income*): cociente entre los pagos mensuales de deuda del prestatario y sus ingresos mensuales totales en el momento de la originación, expresado en porcentaje.
`original_upb`	Numérica continua	Importe original del préstamo en la fecha de firma, redondeado al millar más cercano.
`original_loan_term`	Numérica entera	Plazo original del préstamo en meses, calculado como la diferencia entre la fecha de vencimiento y la fecha del primer pago más uno.
`interest_rate`	Numérica continua	Tipo de interés del préstamo en el momento de la originación, expresado en porcentaje.
`loan_purpose`	Categórica	Finalidad del préstamo: P (compra), C (refinanciación con extracción de capital), N (refinanciación sin extracción de capital) o R (refinanciación no especificada).
`occupancy_status`	Categórica	Uso del inmueble: P (residencia principal), I (inversión) o S (segunda residencia).
`property_type`	Categórica	Tipo de inmueble: SF (vivienda unifamiliar), PU (urbanización planificada), CO (condominio), MH (vivienda prefabricada) o CP (cooperativa).
`channel`	Categórica	Canal de originación del préstamo: R (retail), B (broker), C (correspondent) o T (TPO no especificado).
`first_time_homebuyer_flag`	Binaria	Indica si el prestatario es comprador de vivienda por primera vez: Y (sí) o N (no).
`num_units`	Numérica entera	Número de unidades del inmueble (1 a 4).
`mi_pct`	Numérica entera	Porcentaje de cobertura del seguro hipotecario en el momento de la adquisición por Freddie Mac. Valor 0 indica ausencia de seguro.
`num_borrowers`	Numérica entera	Número de prestatarios obligados al pago del préstamo.

A.1.2 Fichero de *performance*

Cuadro A.2: Variables originales del fichero de *performance* utilizadas en el análisis

Variable	Tipo	Descripción
`loan_sequence_number`	Texto	Identificador único asignado a cada préstamo en el dataset de Freddie Mac. Sigue el formato `PYYQnXXXXXXX`, donde `P` indica el tipo de producto, `YYQn` el año y trimestre de originación, y `XXXXXXX` corresponde a dígitos asignados aleatoriamente.
`monthly_reporting_period`	Fecha (YYYYMM)	Mes de referencia al que corresponde cada registro del fichero de *performance*.
`current_actual_upb`	Numérica continua	Saldo vivo del préstamo al final del mes de *reporting*, incluyendo tanto el principal amortizante como el principal diferido no devengado, si lo hubiera.
`current_loan_delinquency_status`	Numérica entera	Número de cuotas mensuales impagadas, calculado según el método de la Mortgage Bankers Association (MBA). El valor 0 indica que el préstamo está al corriente; valores mayores o iguales a 3 indican situación de default. Toma el valor especial `RA` cuando el préstamo ha sido adjudicado como REO.
`current_interest_rate`	Numérica continua	Tipo de interés vigente en el mes de *reporting*, expresado en porcentaje. Para hipotecas a tipo fijo, permanece constante a lo largo de la vida del préstamo.
`zero_balance_code`	Categórica	Código que identifica el motivo por el que el saldo del préstamo se reduce a cero, es decir, el tipo de evento de terminación. Los valores relevantes para este trabajo son 01 (prepago voluntario), 02 (venta a terceros), 03 (venta corta o *charge-off*), 09 (disposición REO), 15 (venta de nota) y 16 (venta como *reperforming*).
`zero_balance_removal_upb`	Numérica continua	Saldo total pendiente del préstamo inmediatamente antes de la aplicación del código de cierre. Es el único campo de recuperación informado para el código 01 (prepago voluntario).
`net_sale_proceeds`	Numérica continua	Importe remitido a Freddie Mac como resultado de la venta del inmueble o del préstamo, una vez deducidos los gastos de venta de los ingresos brutos.
`mi_recoveries`	Numérica continua	Recuperaciones procedentes del seguro hipotecario (*mortgage insurance*) en caso de pérdida crediticia.
`non_mi_recoveries`	Numérica continua	Recuperaciones no procedentes del seguro hipotecario, incluyendo reembolsos de impuestos o seguros, ingresos por alquiler, saldo positivo en cuentas *escrow* y otros créditos varios.

Continúa en la página siguiente

Variable	Tipo	Descripción
`total_expenses`	Numérica continua	Gastos totales soportados por Freddie Mac en el proceso de adquisición, mantenimiento y venta del inmueble, excluyendo los gastos de venta ya deducidos en `net_sale_proceeds`. Incluye costes legales, mantenimiento y conservación, impuestos y seguros, y gastos varios. En este trabajo se recalcula como el máximo entre el campo original y la suma de sus cuatro componentes individuales.
`delinquent_accrued_interest`	Numérica continua	Importe de los intereses devengados e impagados por el prestatario en el momento del cierre del préstamo. Solo se informa para los códigos de cierre 02, 03, 09 y 15.
`loan_age`	Numérica continua	Número de meses transcurridos desde la originación del préstamo hasta el momento del default.
`default_episode`	Categórica ordinal	Número de episodio de default del prestatario (1 = primer impago, 2 = segundo, etc.). Indica si el prestatario ha incurrido previamente en situaciones de impago.

A.2 Variables construidas

Las siguientes variables no forman parte del dataset original de Freddie Mac, sino que han sido construidas durante el proceso de preparación de datos descrito en el Capítulo 3. Se documentan aquí para facilitar la reproducibilidad del análisis.

Cuadro A.3: Variables construidas durante la preparación de datos

Variable	Tipo	Descripción
`loan_id`	Texto	Identificador único del episodio de default. Se construye concatenando el `loan_sequence_number` original y el número de episodio; por ejemplo, `F05Q10004696_2` identifica el segundo episodio de default del préstamo `F05Q10004696`.
`is_cured`	Binaria (0/1)	Indica si el préstamo ha salido del estado de default en el mes correspondiente. Toma valor 1 cuando acumula al menos tres meses consecutivos con `dq <3` y no se encuentra en estado REO.

Continúa en la página siguiente

Variable	Tipo	Descripción
`terminacion_cura`	Categórica	Clasifica la resolución del episodio como favorable, desfavorable o abierta. Es favorable cuando termina en prepago voluntario, venta como *reperforming* o curación definitiva; desfavorable cuando termina en adjudicación, venta corta, *charge-off* u otros eventos de pérdida; y abierta cuando no existe resolución definitiva al final del periodo de observación.
`resolucion`	Categórica	Indica si el episodio cuenta con una resolución definitiva observada. Toma el valor *completo* cuando existe un código de cierre o una curación confirmada, y *abierto* cuando el episodio sigue activo al final del periodo de observación.
`duration_months`	Numérica entera	Número de meses del episodio desde el primer mes en default, `months_since_default = 0`, hasta el último mes observado. Constituye la variable de tiempo en los modelos de supervivencia. Para los episodios abiertos refleja el tiempo transcurrido hasta el final del periodo de observación.
`EAD`	Numérica continua	Exposición en el momento del default (*Exposure at Default*). Corresponde al valor de `current_actual_upb` en el primer mes del episodio, `months_since_default = 0`.
`cash_flow_t`	Numérica continua	Flujo de caja de recuperación en cada mes del episodio, expresado en valor absoluto. Su construcción varía según el tipo de mes: cero en el mes inicial, diferencia de saldos en meses intermedios, `zero_balance_removal_upb` en prepagos, recuperaciones netas en terminaciones y `current_actual_upb` como flujo artificial en episodios curados sin código de cierre.
`LGD_t`	Numérica continua	*Loss Given Default* calculada en cada mes t del episodio como la fracción del saldo expuesto que no se recupera, descontando los flujos de caja futuros al tipo de interés del primer mes del episodio. El valor en $t = 0$ constituye la LGD definitiva del episodio y es la variable objetivo del modelo.
`weight`	Numérica continua $\in [0, 1]$	Peso asignado a cada observación en el dataset de supervivencia. Se calcula como CF_t/EAD, es decir, la fracción del saldo expuesto recuperada en ese mes. La suma de pesos de cada episodio es igual a 1 por construcción.
`event`	Binaria (0/1)	Indicador de evento en el dataset de supervivencia. Toma valor 1 cuando el flujo de caja del mes es positivo, es decir, cuando hay recuperación, y 0 en caso contrario o en filas de censura.

A.3 Calidad de datos y valores ausentes

El tratamiento de los valores ausentes difiere entre los dos ficheros del dataset por razones metodológicas. En el fichero de originación, los valores ausentes corresponden a información no disponible en el momento de la concesión del préstamo. Por este motivo se aplica un umbral de exclusión: las variables con más del 70 % de valores ausentes se descartan como candidatas al modelo, ya que su escasa cobertura impediría una estimación fiable de su efecto sobre la LGD. Las tres variables excluidas por este criterio se recogen en la Tabla A.4.

Cuadro A.4: Variables excluidas por porcentaje de valores ausentes

Variable	Valores ausentes (%)	Motivo de exclusión
`program_indicator`	97,5	Cobertura insuficiente
`property_valuation_method`	91,8	Cobertura insuficiente
`mi_cancellation_indicator`	77,3	Cobertura insuficiente

En el fichero de *performance*, en cambio, la presencia de valores ausentes en las variables de recuperación no justifica necesariamente su exclusión. Variables como `net_sale_proceeds`, `mi_recoveries`, `non_mi_recoveries`, `delinquent_accrued_interest` o `zero_balance_removal_upb` presentan un alto porcentaje de valores ausentes porque solo se informan en el mes en que se produce el evento de terminación del préstamo, que ocurre una única vez por episodio. Su ausencia mayoritaria es, por tanto, esperada y no indica un problema de calidad, sino la naturaleza puntual de los flujos de recuperación. La Tabla A.5 recoge las variables del fichero de *performance* con más del 70 % de valores ausentes que se conservan en el análisis por esta razón.

No obstante, no todas las variables con alto porcentaje de valores ausentes se conservan en el análisis. Algunas se eliminan porque sus ausencias no responden a la naturaleza puntual de los flujos de recuperación, sino a decisiones previas de depuración de la muestra. En particular, las variables relacionadas con modificaciones contractuales, `modification_cost`, `current_month_modification_cost` y `step_modification_flag`, así como la variable asociada a planes de pago diferido, `deferred_payment_plan`, quedan prácticamente vacías tras excluir los préstamos con estas características durante la construcción de la muestra en el Capítulo 3.

A.4 Análisis univariante

Para evaluar la relación individual de cada variable candidata con la LGD se aplican contrastes no paramétricos sobre las variables utilizadas como covariables en los modelos de supervivencia: la correlación de Spearman para las variables numéricas y el test de Kruskal-Wallis para las variables categóricas. Se consideran significativas las variables con p-valor inferior a 0,01. Aunque el análisis se centra en las variables del fichero de originación, se incluyen también tres variables procedentes del fichero de *performance*, extraídas en el primer mes del episodio ($t = 0$) y tratadas como características fijas del

Cuadro A.5: Variables del fichero de *performance* conservadas pese al porcentaje de valores ausentes

Variable	Valores ausentes (%)	Motivo de conservación
LGD_given	99,93	Solo disponible para préstamos con pérdida real reportada
actual_loss_calculation	99,93	Solo informada en el mes de terminación
defect_settlement_date	99,76	Solo informada en préstamos con defecto
net_sale_proceeds_num	97,08	Solo informada en el mes de terminación
non_mi_recoveries	97,08	Solo informada en el mes de terminación
mi_recoveries	97,08	Solo informada en el mes de terminación
delinquent_accrued_interest	97,07	Solo informada en el mes de terminación
net_sale_proceeds	97,07	Solo informada en el mes de terminación
due_date_of_last_paid_installment	95,81	Solo informada en meses con actividad de pago

préstamo a efectos del modelo: current_interest_rate_ep, el tipo de interés vigente en el momento del default; loan_age, la antigüedad del préstamo en el instante del impago; y default_episode, el número de episodio de default del prestatario.

Cuadro A.6: Correlación de Spearman con la LGD para variables numéricas

Variable	Correlación	p-valor	Significativa
current_interest_rate_ep	0,440	$< 0,001$	Sí
interest_rate	0,440	$< 0,001$	Sí
original_loan_term	0,141	$< 0,001$	Sí
loan_age	−0,114	$< 0,001$	Sí
cltv	0,093	$< 0,001$	Sí
credit_score	−0,091	$< 0,001$	Sí
ltv	0,086	$< 0,001$	Sí
dti	0,073	$< 0,001$	Sí
num_units	0,020	0,001	Sí
mi_pct	0,007	0,230	No
num_borrowers	−0,004	0,542	No
original_upb	−0,001	0,846	No

A.5 Multicolinealidad

El análisis de multicolinealidad se aplica sobre las nueve variables numéricas significativas identificadas en el análisis univariante, utilizando la matriz de correlaciones de Spearman y el Factor de Inflación de la Varianza (VIF). Una correlación absoluta superior a 0,70 entre dos variables indica redundancia informativa relevante, mientras que valores de VIF superiores a 5 sugieren multicolinealidad problemática y valores superiores a 10 indican multicolinealidad severa. Los pares con correlación elevada se recogen en la Tabla A.8, y los valores de VIF por variable se muestran en la Tabla A.9.

Cuadro A.7: Test de Kruskal-Wallis para variables categóricas

Variable	Estadístico	p-valor	Categorías	Significativa
`channel`	1.845,8	$< 0,001$	4	Sí
`postal_region`	627,7	$< 0,001$	9	Sí
`loan_purpose`	299,8	$< 0,001$	3	Sí
`relief_refinance_indicator`	265,1	$< 0,001$	2	Sí
`property_type`	234,8	$< 0,001$	5	Sí
`default_episode`	134,3	$< 0,001$	6	Sí
`occupancy_status`	127,9	$< 0,001$	3	Sí
`super_conforming_flag`	127,5	$< 0,001$	2	Sí
`first_time_homebuyer_flag`	15,2	$< 0,001$	2	Sí
`prepayment_penalty_mortgage_flag`	6,8	0,009	2	Sí

Cuadro A.8: Pares de variables con multicolinealidad detectada

Variable 1	Variable 2	Correlación	Decisión
`current_interest_rate_ep`	`interest_rate`	0,99	Se elimina `interest_rate`.
`cltv`	`ltv`	0,92	Se elimina `cltv`.

Cuadro A.9: Factor de Inflación de la Varianza de las variables numéricas

Variable	VIF	Decisión
`interest_rate`	38,80	Eliminada
`current_interest_rate_ep`	38,57	Conservada
`cltv`	10,11	Eliminada
`ltv`	9,97	Conservada
`original_loan_term`	1,12	Conservada
`credit_score`	1,10	Conservada
`dti`	1,02	Conservada
`loan_age`	1,03	Conservada
`num_units`	1,01	Conservada

A.6 Selección multivariante

La selección *stepwise* hacia adelante parte de las 17 variables candidatas resultantes del análisis de multicolinealidad. Antes de la estimación, las variables categóricas se transforman en variables *dummy*, tomando como categoría de referencia la más frecuente en cada caso. El algoritmo incorpora variables de forma iterativa mientras mejora el AIC, hasta alcanzar un AIC final de 14.036,23 con 24 variables seleccionadas. Sobre el modelo resultante se aplica la corrección de Bonferroni, con un umbral de significancia ajustado de 0,00208 $(0,05/24)$. Las variables que no superan este umbral son `postal_region_6`, `default_episode_3`, `channel_R`, `dti`, `num_units` y `property_type_CP`. La Tabla A.10 recoge los coeficientes y p-valores del modelo, indicando qué variables superan el umbral corregido.

Cuadro A.10: Resultados del modelo *stepwise* AIC con corrección de Bonferroni

Variable	Coeficiente	p-valor	Sig. Bonferroni
`current_interest_rate_ep`	0,746	$< 0,001$	Sí
`loan_age`	−0,285	$< 0,001$	Sí
`ltv`	0,234	$< 0,001$	Sí
`original_loan_term`	0,096	$< 0,001$	Sí
`credit_score`	0,070	0,002	Sí
`postal_region_3`	0,077	$< 0,001$	Sí
`postal_region_4`	0,064	$< 0,001$	Sí
`postal_region_8`	0,057	$< 0,001$	Sí
`postal_region_2`	−0,032	$< 0,001$	Sí
`first_time_homebuyer_flag_Y`	0,025	$< 0,001$	Sí
`channel_T`	0,010	$< 0,001$	Sí
`loan_purpose_N`	−0,029	$< 0,001$	Sí
`loan_purpose_P`	−0,083	$< 0,001$	Sí
`occupancy_status_P`	−0,049	$< 0,001$	Sí
`default_episode_2`	−0,064	$< 0,001$	Sí
`property_type_PU`	−0,055	$< 0,001$	Sí
`property_type_SF`	−0,039	$< 0,001$	Sí
`postal_region_7`	−0,064	$< 0,001$	Sí
`postal_region_6`	0,041	$< 0,001$	No
`default_episode_3`	0,010	0,034	Sí
`channel_R`	0,047	0,058	No
`dti`	0,028	0,077	No
`num_units`	−0,018	0,105	No
`property_type_CP`	−0,078	0,105	No

A.7 Weight of Evidence e Information Value

La siguiente tabla recoge los resultados del análisis de WOE e IV para las variables numéricas discretizadas. La Tabla A.11 muestra el IV agregado por variable, que determina su

poder discriminante global sobre la LGD. La discretización de cada variable se obtuvo tras explorar distintas combinaciones de tramos manualmente, seleccionando aquella segmentación que mejor cumplía los criterios descritos en el apartado 3.5 y maximizaba simultáneamente el IV resultante.

Las variables `credit_score` y `dti` presentan un IV inferior a 0,02, lo que indica ausencia de poder predictivo sobre la LGD. Aunque se realizaron pruebas incorporándolas al modelo de Cox PH, su inclusión no mejoró la capacidad predictiva del modelo e introdujo ruido en las estimaciones, por lo que fueron finalmente descartadas.

Cuadro A.11: Resumen de IV por variable numérica

Variable	IV	Poder discriminante	Decisión
`current_interest_rate_ep`	0,220	Moderado	Conservada
`ltv`	0,041	Débil	Conservada
`loan_age`	0,046	Débil	Conservada
`original_loan_term`	0,027	Débil	Conservada
`credit_score`	0,010	Sin poder	Descartada
`dti`	0,008	Sin poder	Descartada

A.8 Distribución por categorías de las variables finales seleccionadas

La Tabla A.12 recoge la distribución por tramos o categorías de las variables finales seleccionadas para la modelización.

Cuadro A.12: Distribución por categorías de las variables finales seleccionadas

Variable	Tipo	Tramo o categoría	N	%
`current_interest_rate_ep`	Numérica	$< 4\,\%$	3.315	11,6 %
`current_interest_rate_ep`	Numérica	4–5 %	5.634	19,8 %
`current_interest_rate_ep`	Numérica	5–6 %	7.113	25,0 %
`current_interest_rate_ep`	Numérica	6–7 %	10.808	37,9 %
`current_interest_rate_ep`	Numérica	$> 7\,\%$	1.632	5,7 %
`ltv`	Numérica	$< 60\,\%$	3.964	13,9 %
`ltv`	Numérica	60–80 %	15.330	53,8 %
`ltv`	Numérica	80–90 %	3.792	13,3 %

Continúa en la página siguiente

Variable	Tipo	Tramo o categoría	N	%
`ltv`	Numérica	90–100 %	4.388	15,4 %
`ltv`	Numérica	> 100 %	1.028	3,6 %
`original_loan_term`	Numérica	≤ 15 años	2.157	7,6 %
`original_loan_term`	Numérica	16–30 años	26.306	92,3 %
`original_loan_term`	Numérica	> 30 años	39	0,1 %
`loan_purpose`	Categórica	C (refinanciación con extracción)	9.319	32,7 %
`loan_purpose`	Categórica	N (refinanciación sin extracción)	8.250	29,0 %
`loan_purpose`	Categórica	P (compra)	10.933	38,3 %
`occupancy_status`	Categórica	I (inversión)	2.285	8,0 %
`occupancy_status`	Categórica	P (residencia principal)	25.082	88,0 %
`occupancy_status`	Categórica	S (segunda residencia)	1.135	4,0 %
`property_type`	Categórica	CO (condominio)	2.702	9,5 %
`property_type`	Categórica	CP (cooperativa)	52	0,2 %
`property_type`	Categórica	MH (vivienda prefabricada)	450	1,6 %
`property_type`	Categórica	PU (urbanización planificada)	4.566	16,0 %
`property_type`	Categórica	SF (vivienda unifamiliar)	20.732	72,7 %
`first_time_homebuyer_flag`	Categórica	N	24.350	85,4 %
`first_time_homebuyer_flag`	Categórica	Y	4.152	14,6 %
`postal_region`	Categórica	1	2.167	7,6 %
`postal_region`	Categórica	2	2.946	10,3 %
`postal_region`	Categórica	3	5.496	19,3 %
`postal_region`	Categórica	4	3.290	11,5 %
`postal_region`	Categórica	5	1.531	5,4 %
`postal_region`	Categórica	6	2.871	10,1 %
`postal_region`	Categórica	7	2.726	9,6 %
`postal_region`	Categórica	8	3.364	11,8 %
`postal_region`	Categórica	9	4.111	14,4 %

Apéndice B

Anexo B: resultados del entrenamiento y validación de los modelos

Este anexo recoge el detalle complementario del proceso de entrenamiento y selección de hiperparámetros descrito en la Sección 4.4 e información complementaria a los resultados de los modelos del capítulo 5.

B.1 Modelo de regresión base

Como paso previo a los modelos de análisis de supervivencia, se estimó un modelo de regresión OLS utilizando las variables finales seleccionadas en el Capítulo 3. Además de los efectos principales, se probaron términos de interacción para comprobar si el efecto de algunas variables cambiaba en función de otras características del préstamo. La Tabla B.1 muestra los coeficientes finales del modelo OLS. La mayoría de los predictores incluidos presentan un p-valor inferior a 0,05; las variables que no alcanzaron significatividad estadística en la especificación final fueron `loan_age_ord`, `property_type_CP`, `postal_region_2`, `postal_region_9` y `rate_x_ltv`. Cabe señalar que `loan_age_ord` sí resultaba significativa de forma individual ($p < 0,001$) antes de introducir los términos de interacción, pero pierde significatividad al incluir `rate_x_loanage`, que recoge parte de su efecto sobre la LGD.

Cuadro B.1: Coeficientes finales del modelo OLS

Variable	Coef.	Error típ.	t	p-valor
const	0,084	0,015	5,609	$< 0,001$
current_interest_rate_ep_ord	0,556	0,015	37,451	$< 0,001$
ltv_ord	0,075	0,015	4,967	$< 0,001$
original_loan_term_ord	0,071	0,016	4,524	$< 0,001$
first_time_homebuyer_flag_Y	$-0,014$	0,007	$-2,000$	0,045
occupancy_status_P	$-0,061$	0,006	$-9,798$	$< 0,001$
property_type_PU	$-0,085$	0,008	$-10,781$	$< 0,001$
property_type_SF	$-0,049$	0,006	$-7,655$	$< 0,001$
loan_purpose_P	$-0,072$	0,006	$-12,681$	$< 0,001$
loan_purpose_N	$-0,028$	0,005	$-5,184$	$< 0,001$
postal_region_3	0,067	0,009	7,822	$< 0,001$
postal_region_4	0,066	0,009	7,146	$< 0,001$
postal_region_5	0,029	0,011	2,663	0,008
postal_region_6	0,051	0,009	5,356	$< 0,001$
postal_region_7	$-0,057$	0,010	$-5,852$	$< 0,001$
postal_region_8	0,050	0,009	5,389	$< 0,001$
postal_region_9	0,016	0,009	1,847	0,065
default_episode_2	$-0,078$	0,009	$-8,509$	$< 0,001$
rate_x_loanage	$-0,276$	0,022	$-12,543$	$< 0,001$

B.2 Modelo Cox PH

En el modelo Cox PH se evaluó el penalizador de regularización, manteniendo constante el resto de la especificación del modelo. La Tabla B.2 resume el rango analizado y el valor seleccionado para el entrenamiento final.

Cuadro B.2: Resultados de validación cruzada para el penalizador del modelo Cox PH

Penalizador λ	C medio	Desv. típica	Seleccionado
0,001	0,5821	0,0032	No
0,010	0,5821	0,0032	Sí
0,050	0,5819	0,0029	No
0,100	0,5814	0,0028	No
0,500	0,5793	0,0025	No

El valor $\lambda = 0{,}01$ se selecciona por obtener el mayor índice de concordancia medio en validación cruzada, aunque las diferencias entre penalizadores fueron reducidas. Esto indica que el modelo no requiere una regularización fuerte para estabilizar la estimación.

En cuanto al modelo construido, la Tabla B.3 recoge los coeficientes del modelo Cox PH para las variables significativas al 1 %. Se muestra el coeficiente estimado, su exponencial o *hazard ratio*, y el p-valor asociado. Un *hazard ratio* superior a 1 indica que la variable aumenta la velocidad de recuperación del préstamo, mientras que un valor inferior a 1 indica el efecto contrario, asociándose por tanto a una mayor LGD esperada.

Cuadro B.3: Coeficientes significativos del modelo Cox PH

Variable	Coef.	Hazard ratio	p-valor
`current_interest_rate_ep_ord`	$-1,2253$	$0,2937$	$< 0,001$
`ltv_ord`	$-0,2199$	$0,8026$	$< 0,001$
`original_loan_term_ord`	$-0,4029$	$0,6684$	$< 0,001$
`loan_age_ord`	$0,3696$	$1,4471$	$< 0,001$
`property_type_PU`	$0,1939$	$1,2139$	$< 0,001$
`loan_purpose_P`	$0,2191$	$1,2450$	$< 0,001$
`loan_purpose_N`	$0,1139$	$1,1206$	$< 0,001$
`postal_region_2`	$0,1505$	$1,1624$	$< 0,001$
`postal_region_5`	$0,0941$	$1,0987$	$0,007$
`postal_region_7`	$0,2564$	$1,2923$	$< 0,001$
`postal_region_8`	$0,1679$	$1,1828$	$< 0,001$
`postal_region_9`	$0,1836$	$1,2015$	$< 0,001$
`default_episode_2`	$0,1152$	$1,1220$	$< 0,001$

B.3 *Random Survival Forest*

Para el *Random Survival Forest* se evaluaron 16 combinaciones de hiperparámetros mediante validación cruzada. Los parámetros analizados controlan la complejidad del bosque y de los árboles individuales La Tabla B.4 muestra las configuraciones más representativas evaluadas.

Cuadro B.4: Resultados de validación cruzada para el *Random Survival Forest*

Número de árboles	Mín. obs. nodo	Variables por split	C medio	Desv. típica
100	10	$\sqrt{p}$	0,5841	0,0044
200	10	$\sqrt{p}$	0,5845	0,0043
300	10	$\sqrt{p}$	0,5846	0,0044
200	20	$0{,}5p$	0,5876	0,0032
200	15	$0{,}5p$	0,5868	0,0029

La configuración finalmente empleada para entrenar el RSF corresponde a 200 árboles, 20 observaciones mínimas por nodo terminal y $0{,}5p$ variables candidatas por división, al presentar el mayor índice de concordancia medio en la validación cruzada interna. Este criterio permite seleccionar el modelo con mejor capacidad de ordenación fuera de muestra, manteniendo el conjunto de test reservado para la evaluación final.

Una vez ajustado el modelo, se calcula la importancia de cada variable (VIMP) mediante permutación sobre el conjunto de validación: para cada variable, se permutan aleatoriamente sus valores y se mide cuánto empeora el índice de concordancia del modelo. Una mayor caída en el rendimiento indica una mayor relevancia de la variable para las predicciones del modelo. La Tabla B.5 recoge las variables con importancia igual o superior a 0,001.

Cuadro B.5: Importancia de variables del *Random Survival Forest*

Variable	Importancia	Desv. típica
`current_interest_rate_ep_ord`	0,0392	0,0064
`loan_age_ord`	0,0083	0,0009
`postal_region_7`	0,0034	0,0007
`original_loan_term_ord`	0,0021	0,0004
`property_type_PU`	0,0018	0,0008
`loan_purpose_P`	0,0017	0,0003
`channel_C`	0,0011	0,0002

Bibliografía

[1] Jiří Witzany, Michal Rychnovský y Pavel Charamza. *Survival Analysis in LGD Modeling*. IES Working Paper 2/2010. Prague: Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague, 2010.

[2] Basel Committee on Banking Supervision. *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Technical report. Basel: Bank for International Settlements, 2006.

[3] Edward L. Kaplan y Paul Meier. "Nonparametric Estimation from Incomplete Observations". En: *Journal of the American Statistical Association* 53.282 (1958), págs. 457-481. DOI: 10.1080/01621459.1958.10501452.

[4] David R. Cox. "Regression Models and Life-Tables". En: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), págs. 187-202. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

[5] Leo Breiman et al. *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[6] Leo Breiman. "Random Forests". En: *Machine Learning* 45.1 (2001), págs. 5-32. DOI: 10.1023/A:1010933404324.

[7] Hemant Ishwaran et al. "Random Survival Forests". En: *The Annals of Applied Statistics* 2.3 (2008), págs. 841-860. DOI: 10.1214/08-AOAS169.

[8] Wayne Nelson. "Theory and Applications of Hazard Plotting for Censored Failure Data". En: *Technometrics* 14.4 (1972), págs. 945-966. DOI: 10.1080/00401706.1972.10488991.

[9] Odd O. Aalen. "Nonparametric Inference for a Family of Counting Processes". En: *The Annals of Statistics* 6.4 (1978), págs. 701-726. DOI: 10.1214/aos/1176344247.

[10] Freddie Mac. *Single-Family Loan-Level Dataset: General User Guide*. User guide. Federal Home Loan Mortgage Corporation, 2026.

[11] Frank E. Harrell et al. "Evaluating the Yield of Medical Tests". En: *JAMA* 247.18 (1982), págs. 2543-2546. DOI: 10.1001/jama.1982.03320430047030.