



ICADE, Facultad de Ciencias Económicas y Empresariales,
Universidad Pontificia de Comillas

Programa: Doble Grado en Administración de Empresas y Business Analytics

**Modelo Predictivo de Riesgo de Pérdida de Contratos de Arrendamiento en
Infraestructuras de Telecomunicaciones**

Autor: Carlos Mijangos Ridruejo

Supervisor: Rafael Castellote Azorín

17 de junio 2026

Resumen

El presente Trabajo de Fin de Grado desarrolla y valida empíricamente el modelo predictivo que constituye el núcleo de SiteGuard, una solución analítica destinada al sector de las towercos europeas. Partiendo de un conjunto de datos real de 11.501 eventos de renovación de contratos de arrendamiento de suelo cedido por una towerco europea, se construye un modelo de clasificación binaria capaz de asignar a cada evento de renovación una probabilidad cuantitativa de pérdida y de generar, a partir de ella, un ranking de riesgo accionable para los equipos de gestión de cartera.

La aproximación metodológica prioriza la interpretabilidad sobre la complejidad. Tras un análisis exploratorio que identifica las señales operativas y contractuales asociadas a la pérdida, se entrena una regresión logística con regularización L2 y ponderación de clases para gestionar el desbalanceo del problema (78,6% de continuidades frente a 21,4% de pérdidas). El modelo final alcanza un área bajo la curva ROC de 0,913, con una capacidad de recuperación de la clase minoritaria del 83%, y demuestra una eficacia de priorización elevada: el 20% de eventos de mayor riesgo concentra cerca del 68% de las pérdidas reales, lo que multiplica por más de tres la eficiencia de una revisión aleatoria de la cartera.

Los resultados confirman la viabilidad técnica de la propuesta y aportan el respaldo empírico sobre el que se construye el plan de negocio desarrollado en el Trabajo de Fin de Grado paralelo de la titulación de Administración y Dirección de Empresas.

Palabras clave: predicción de churn, clasificación binaria, regresión logística, towerco, contratos de arrendamiento, aprendizaje automático, clases desbalanceadas, ranking de riesgo.

Abstract

This Undergraduate Work develops and empirically validates the predictive model at the core of SiteGuard, an analytical solution designed for the European towerco sector. Drawing on a real dataset of 11,501 ground-lease renewal events provided by a European towerco, a binary classification model is built to assign each renewal event a quantitative probability of loss and to generate, from it, an actionable risk ranking for portfolio management teams.

The methodological approach prioritizes interpretability over complexity. Following an exploratory analysis that identifies the operational and contractual signals associated with loss, a logistic regression with L2 regularization and class weighting is trained to handle the problem's imbalance (78.6% continuity versus 21.4% loss). The final model achieves an area under the ROC curve of 0.913, with an 83% recall of the minority class, and demonstrates strong prioritization power: the top 20% of highest-risk events concentrates close to 68% of actual losses, more than tripling the efficiency of a random portfolio review.

The results confirm the technical feasibility of the proposal and provide the empirical backbone supporting the business plan developed in the parallel Undergraduate Work of the Business Administration degree program.

Keywords: churn prediction, binary classification, logistic regression, towerco, lease contracts, machine learning, imbalanced classes, risk ranking.

Índice

1.	Introducción	5
1.1.	Contexto y motivación.....	5
1.2.	Objetivos del trabajo y pregunta de investigación.....	6
1.3.	Estructura del documento	7
2.	Revisión de literatura	8
2.1.	La predicción de abandono: evolución metodológica y estado del arte	8
2.2.	Aplicaciones en infraestructuras y mercados B2B: el gap identificado	8
2.3.	Técnicas de clasificación binaria	9
2.4.	Métricas de evaluación en problemas con clases desbalanceadas	11
3.	Datos y metodología	12
3.1.	Descripción del conjunto de datos	12
3.2.	Preprocesamiento de datos.....	14
3.3.	Selección de variables y tratamiento de la multicolinealidad.....	16
3.4.	Estrategia de validación y gestión del desbalanceo	17
3.5.	Modelos y criterios de evaluación	17
4.	Resultados	19
4.1.	Análisis exploratorio y estadística descriptiva.....	19
4.2.	Comparativa de modelos.....	22
4.3.	Interpretación del modelo final.....	25
4.4.	Aplicación práctica: el ranking de riesgo.....	27
5.	Discusión.....	30
5.1.	Interpretación de los resultados y conexión con el negocio	30
5.2.	Implicaciones para el negocio.....	31
5.3.	Comparación con la literatura.....	32
5.4.	Limitaciones y líneas de mejora	33
6.	Conclusiones	35
7.	Declaración del uso de la inteligencia artificial.....	38
8.	Bibliografía	40

1. Introducción

1.1. Contexto y motivación

El sector de las infraestructuras pasivas de telecomunicaciones europeo ha experimentado en la última década una transformación estructural profunda. La progresiva externalización por parte de los operadores de redes móviles de la propiedad y gestión de sus torres, mástiles y emplazamientos ha consolidado un mercado mayorista dominado por operadores especializados, las towercos, cuya cuota de mercado en Europa pasó del 13% en 2014 al 39% en 2023 (EY-Parthenon, 2024). Este modelo descansa sobre un activo aparentemente intangible pero económicamente decisivo: el contrato de arrendamiento del suelo, azotea o estructura sobre la que se instala la infraestructura.

La dimensión del mercado ibérico ilustra la magnitud del problema. España cuenta con un parque aproximado de 37.000 emplazamientos y Portugal con unos 14.000, lo que sitúa la península en torno a 51.000 torres gestionadas por un número reducido de operadores (EY-Parthenon, 2024). El operador dominante, Cellnex Telecom, gestionaba a finales de 2025 cerca de 8.863 emplazamientos en España, seguido por Vantage Towers, TOTEM España y operadores regionales como Axió (Cellnex Telecom, 2025; Mordor Intelligence, 2025). Se trata, por tanto, de un mercado maduro y concentrado, en el que cada operador administra carteras de miles de contratos cuya conservación resulta crítica para la estabilidad de sus ingresos recurrentes.

Estos contratos, firmados habitualmente por plazos iniciales de alrededor de veinte años, se estructuran en eventos de renovación periódicos. Cada uno de estos eventos constituye una decisión independiente del propietario: salvo que comunique formalmente lo contrario dentro del plazo estipulado, el contrato se renueva de forma automática. Existe, por tanto, una ventana de oportunidad acotada en la que la towerco puede intervenir para preservar el emplazamiento antes de que la decisión se materialice. La pérdida de un emplazamiento no es un suceso menor: la literatura sectorial sitúa el coste de desmantelaje físico de una instalación en un rango de entre 55.000 y 165.000 euros, al que se añaden la pérdida de cobertura, los costes de búsqueda y negociación de un nuevo emplazamiento y los efectos reputacionales sobre los operadores alojados (Steel In The Air, 2025; KPR Consult/Capacity, 2015).

Esta exposición económica adquiere mayor relevancia en la fase actual del sector. Tras una década de crecimiento basado en adquisiciones, las towercos europeas han girado hacia un modelo de crecimiento orgánico que prioriza la rentabilización de la cartera existente sobre la

expansión (S&P Global, 2025). En este contexto, la conservación de cada contrato se convierte en una palanca de creación de valor tan importante como la captación de nuevos emplazamientos, lo que eleva el interés estratégico de cualquier herramienta capaz de anticipar y prevenir las pérdidas.

Pese a la magnitud del problema, el sector no dispone hoy de herramientas sistemáticas que permitan anticipar qué eventos de renovación se encuentran en riesgo. La gestión de cartera se apoya en buena medida en el conocimiento tácito de los gestores y en señales fragmentadas (incidencias, reclamaciones, retrasos en pagos) que rara vez se consolidan con la antelación necesaria. El resultado es una pérdida silenciosa de activos que erosiona los ingresos recurrentes y el valor de la cartera. La motivación de este trabajo nace precisamente de esa brecha: aplicar las técnicas de aprendizaje automático maduras en la predicción de abandono de clientes a un problema económicamente relevante que no ha recibido atención específica en la literatura.

El presente trabajo se sitúa en la intersección entre la analítica de datos y la gestión estratégica de activos de infraestructuras, proponiendo un enfoque predictivo que permita anticipar qué contratos presentan mayor probabilidad de pérdida y, en consecuencia, priorizar las acciones de retención sobre aquellos emplazamientos donde el riesgo es más elevado.

1.2. Objetivos del trabajo y pregunta de investigación

El objetivo general de este Trabajo de Fin de Grado es desarrollar y validar empíricamente un modelo predictivo capaz de estimar, para cada evento de renovación de un contrato de arrendamiento de una towerco, la probabilidad de que el propietario decida no renovar, generando un instrumento de priorización útil para la gestión de cartera. La pregunta de investigación que articula el trabajo puede formularse así: ¿es posible anticipar, a partir de las características contractuales, físicas y operativas registradas por la towerco, qué eventos de renovación presentan un riesgo elevado de pérdida, con una precisión suficiente para orientar la acción comercial?

De este objetivo general se derivan cuatro objetivos específicos. En primer lugar, caracterizar el conjunto de datos y depurar las variables disponibles, identificando posibles fuentes de fuga de información y de sesgo. En segundo lugar, identificar mediante análisis exploratorio las variables asociadas a la pérdida de un emplazamiento. En tercer lugar, entrenar y evaluar un modelo de clasificación binaria interpretable, gestionando adecuadamente el desbalanceo de la variable objetivo. Y, en cuarto lugar, traducir la salida del modelo en un

ranking de riesgo accionable y evaluar su eficacia de priorización, que es el output que el producto SiteGuard entrega al cliente.

Conviene precisar el encaje de este trabajo dentro de un proyecto más amplio. El modelo aquí desarrollado constituye el núcleo predictivo de SiteGuard, cuya estrategia comercial y plan de negocio se desarrollan en un Trabajo de Fin de Grado paralelo en la titulación de Administración y Dirección de Empresas (Mijangos, 2026). Ambos trabajos son complementarios: este aporta la validación técnica de la solución, y aquel construye la empresa que la lleva al mercado.

1.3. Estructura del documento

El documento se organiza en seis secciones. Tras esta introducción, la sección 2 revisa la literatura sobre predicción de abandono y técnicas de clasificación binaria, situando la aportación del trabajo. La sección 3 describe los datos y la metodología, incluyendo el preprocesamiento, la estrategia de validación y los modelos seleccionados. La sección 4 presenta los resultados del análisis exploratorio, la comparativa de modelos, la interpretación del modelo final y su aplicación práctica como ranking de riesgo. La sección 5 discute los resultados y sus limitaciones, y la sección 6 recoge las conclusiones. Cierran el documento la declaración del uso de inteligencia artificial y la bibliografía.

2. Revisión de literatura

2.1. La predicción de abandono: evolución metodológica y estado del arte

La predicción de abandono de clientes, conocida en la literatura anglosajona como churn prediction, constituye uno de los campos de aplicación más consolidados del aprendizaje automático en el sector de las telecomunicaciones. La lógica económica subyacente es directa: dado que el coste de adquirir un nuevo cliente supera ampliamente el de retener uno existente, identificar anticipadamente a los clientes con mayor probabilidad de abandono permite focalizar las acciones de retención y optimizar la asignación de recursos (Ahmad et al., 2019).

La literatura documenta una evolución progresiva en las técnicas empleadas. Los primeros trabajos se apoyaban en modelos estadísticos clásicos, principalmente la regresión logística y los árboles de decisión, valorados por su transparencia y su facilidad de interpretación. Un árbol de decisión segmenta la población mediante reglas sucesivas sobre las variables (por ejemplo, si la antigüedad supera un umbral o si el número de incidencias excede cierto valor), generando grupos con distinta propensión al abandono. Posteriormente, el campo evolucionó hacia los métodos de ensamble, que combinan múltiples árboles para mejorar la capacidad predictiva: el bosque aleatorio (random forest) promedia árboles entrenados sobre muestras distintas, mientras que el gradient boosting los construye de forma secuencial, corrigiendo en cada paso los errores del anterior. Más recientemente, las redes neuronales profundas han incorporado la capacidad de modelar relaciones altamente no lineales a costa de una mayor opacidad.

Los resultados empíricos confirman la madurez del campo. En un trabajo de referencia sobre el operador SyriaTel, Ahmad et al. (2019) alcanzan un área bajo la curva ROC del 93,3% combinando gradient boosting con variables derivadas del análisis de redes sociales del cliente. Chang et al. (2024) y Zhang (2025) documentan resultados comparables en otros contextos de aplicación. Esta evolución refleja una tensión recurrente en la disciplina: la mejora del rendimiento predictivo que ofrecen los modelos más complejos se obtiene con frecuencia a costa de la interpretabilidad, un atributo especialmente valorado cuando el modelo debe integrarse en procesos de decisión empresariales donde el responsable necesita comprender y justificar cada predicción.

2.2. Aplicaciones en infraestructuras y mercados B2B: el gap identificado

La literatura sobre predicción de abandono se ha concentrado de forma abrumadora en el cliente final de mercados masivos, particularmente en retail, servicios financieros y, en menor

medida, en software como servicio dirigido a empresas. La aplicación de estas técnicas a la retención de contratos de arrendamiento de infraestructura, donde el cliente cuya decisión se predice es un propietario de suelo o azotea con motivaciones muy distintas a las de un consumidor de telefonía, no ha recibido atención comparable.

Esta brecha entre la madurez metodológica del campo y su aplicación a un problema económicamente relevante para un sector en crecimiento define el espacio de innovación que este trabajo ocupa. El problema presenta además particularidades propias que justifican un tratamiento específico: la decisión del propietario es discreta y periódica (ligada a eventos de renovación), el desenlace está fuertemente desbalanceado, y las señales disponibles combinan variables contractuales estables con variables operativas dinámicas como el historial de incidencias. Estas características aproximan el problema a la predicción de abandono contractual más que al abandono voluntario continuo característico del retail. Adicionalmente, mientras que en el churn de consumo el comportamiento del cliente genera múltiples señales de uso continuas, en el caso de un propietario de suelo las señales son más escasas y de naturaleza fundamentalmente contractual y relacional, lo que confiere un valor particular a las incidencias registradas como proxies de la satisfacción del propietario.

2.3. Técnicas de clasificación binaria

La clasificación binaria agrupa el conjunto de técnicas orientadas a asignar cada observación a una de dos clases mutuamente excluyentes, en este caso continuidad o pérdida. Entre los modelos disponibles, la regresión logística ocupa una posición central por su solidez teórica y su interpretabilidad. El modelo estima la probabilidad de pertenencia a la clase positiva mediante la función logística o sigmoide, que transforma cualquier combinación lineal de las variables explicativas en un valor acotado entre cero y uno, interpretable como una probabilidad (LaValley, 2008).

Formalmente, si se denota por z la combinación lineal de las variables explicativas ponderadas por sus coeficientes, la probabilidad estimada de pérdida se obtiene como $P(y=1) = 1 / (1 + e^{(-z)})$, donde $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$. La figura 2.1 representa esta función, cuya forma en S garantiza que las predicciones se mantengan siempre en el rango de probabilidad válido. La principal virtud del modelo reside en la interpretabilidad de sus coeficientes: el exponencial de cada coeficiente, conocido como odds ratio, indica en qué factor se multiplica la razón de probabilidades de pérdida ante un incremento unitario de la variable correspondiente, manteniendo constantes las demás (Hosmer et al., 2013). Un odds ratio

superior a uno señala una variable que incrementa el riesgo, y uno inferior a la unidad, una variable protectora.

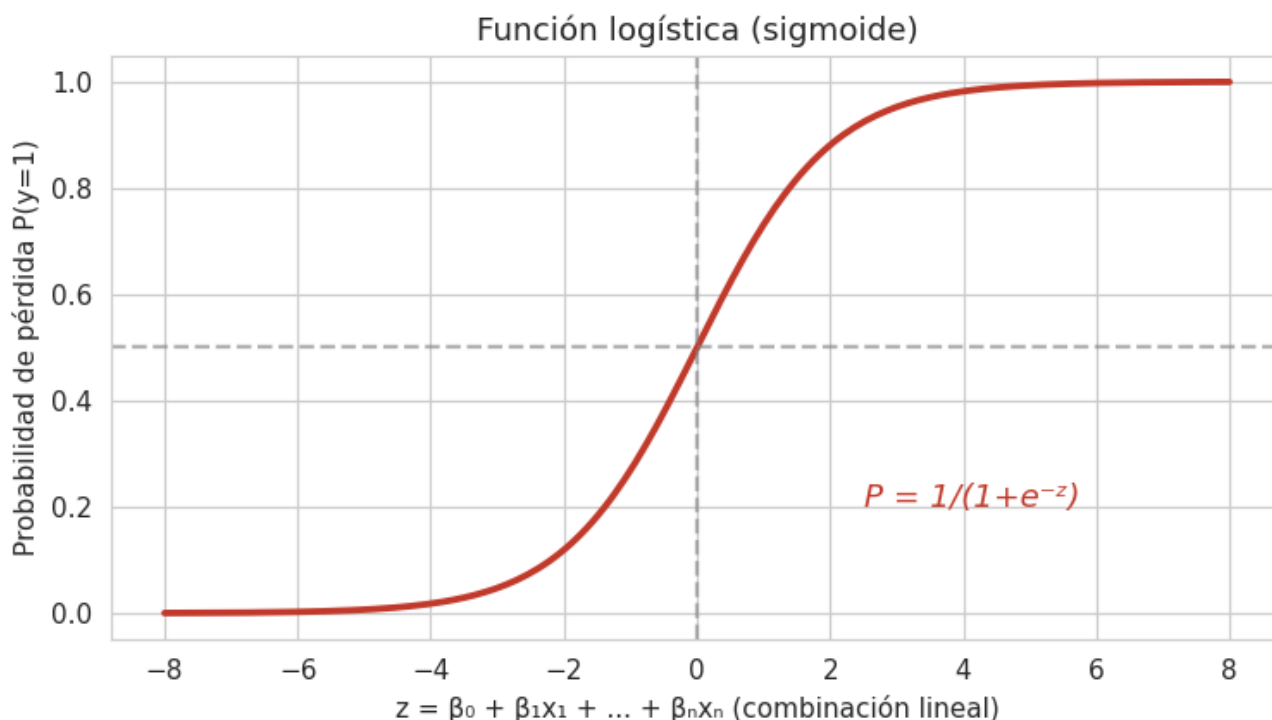


Figura 2.1. Función logística (sigmoide) que sustenta la regresión logística. Elaboración propia.

Frente a la regresión logística, los métodos de ensamble basados en árboles, como el random forest o el gradient boosting, construyen combinaciones de múltiples árboles de decisión capaces de capturar relaciones no lineales e interacciones complejas de forma automática. Estos métodos suelen ofrecer un rendimiento predictivo superior, pero su lógica interna es considerablemente más opaca, lo que dificulta la justificación de cada predicción ante un usuario de negocio. La elección entre ambos paradigmas no es trivial y depende del contexto de aplicación: en entornos donde la transparencia metodológica y la capacidad de explicar cada predicción son requisitos (como ocurre en la gestión de cartera de una towerco, donde el gestor necesita entender por qué un contrato concreto está en riesgo para diseñar la acción correctiva) la regresión logística constituye una opción especialmente adecuada.

Este trabajo adopta deliberadamente la regresión logística como modelo central. La decisión responde a dos consideraciones. En primer lugar, la interpretabilidad: la salida no es una caja negra, sino una probabilidad acompañada de las variables que más contribuyen al riesgo de cada caso. En segundo lugar, la robustez y simplicidad, que facilitan tanto la implantación como la defensa metodológica del modelo.

2.4. Métricas de evaluación en problemas con clases desbalanceadas

La evaluación de un clasificador entrenado sobre un conjunto de datos desbalanceado exige un cuidado particular en la elección de las métricas. La exactitud o accuracy, definida como la proporción de aciertos sobre el total, resulta engañosa en estos contextos: en un problema con un 78,6% de continuidades, un clasificador que prediga siempre continuidad alcanzaría una exactitud del 78,6% sin identificar una sola pérdida. Por ello, la literatura recomienda apoyarse en métricas que distingan el comportamiento del modelo sobre la clase minoritaria (He y Garcia, 2009).

Las métricas empleadas en este trabajo se definen a partir de los cuatro resultados posibles de una clasificación binaria: verdaderos positivos (pérdidas correctamente identificadas), falsos positivos (continuidades señaladas erróneamente como pérdida), verdaderos negativos (continuidades correctamente identificadas) y falsos negativos (pérdidas no detectadas). Sobre esta base, la precisión sobre la clase pérdida mide qué proporción de los eventos señalados como riesgo resultan efectivamente pérdidas; la sensibilidad o recall mide qué proporción de las pérdidas reales logra detectar el modelo; y el F1, su media armónica, equilibra ambas. El área bajo la curva ROC (AUC-ROC) resume la capacidad de discriminación del modelo con independencia del umbral de decisión, y el área bajo la curva de precisión-sensibilidad (PR-AUC) ofrece una visión más informativa cuando la clase positiva es minoritaria. Por último, el pseudo- R^2 de McFadden ofrece una medida del ajuste global del modelo análoga al coeficiente de determinación de la regresión lineal, basada en la comparación de la verosimilitud del modelo con la de un modelo nulo.

Dada la asimetría de costes del problema (no detectar una pérdida es mucho más costoso que revisar innecesariamente un contrato que se habría renovado) se concede prioridad analítica a la sensibilidad sobre la clase pérdida y a la capacidad de priorización del ranking resultante.

3. Datos y metodología

3.1. Descripción del conjunto de datos

El conjunto de datos utilizado procede de una towerco europea y recoge 11.501 eventos de renovación de contratos de arrendamiento, descritos mediante 71 variables originales. La unidad de análisis es el evento de renovación, no el contrato completo: un mismo contrato puede generar varios eventos a lo largo de su vida (el dataset contiene 4.807 contratos y 4.233 emplazamientos únicos), y cada evento constituye una decisión independiente del propietario. Las variables se organizan en tres bloques temáticos: contractual, del emplazamiento y operativo.

El bloque contractual recoge las características legales y económicas del contrato. El bloque del emplazamiento describe las características físicas y jurídicas del activo. Y el bloque operativo sintetiza el historial de incidencias o tickets asociado a cada contrato, organizado a su vez en cuatro subgrupos: el volumen general de incidencias, su distribución por categoría, su criticidad y la satisfacción registrada en su resolución. La tabla 3.1 presenta, a modo de diccionario, una tabla de las variables por bloque empleadas en el modelo.

Variable	Bloque	Tipo	Descripción
Duración contrato	Contractual	Numérica	Duración pactada del contrato (días)
N_renew	Contractual	Numérica	Nº de renovaciones previas del contrato
N_renta	Contractual	Numérica	Nº de revisiones de renta registradas
begin_contract_days	Contractual	Numérica	Antigüedad del contrato desde su firma (días)
Last_annual_Rent	Contractual	Numérica	Renta anual del contrato (€)
P_rent_last_change	Contractual	Numérica	% de variación en la última revisión de renta
last_renewal_days	Contractual	Numérica	Días transcurridos desde la última renovación
N_own	Emplazamiento	Numérica	Nº de propietarios del emplazamiento
Superficie Contratada	Emplazamiento	Numérica	Superficie de suelo o azotea contratada (m²)
num_tickets	Operativo	Numérica	Nº total de incidencias (tickets) registradas
N_tickets_still_open	Operativo	Numérica	Nº de incidencias actualmente abiertas
avg_ticket_duration_days	Operativo	Numérica	Duración media de resolución de incidencias (días)

Variable	Bloque	Tipo	Descripción
N_cat_Finance	Operativo	Numérica	Nº de incidencias de categoría financiera
N_cat_Legal	Operativo	Numérica	Nº de incidencias de categoría legal
N_cat_Administrative	Operativo	Numérica	Nº de incidencias de categoría administrativa
N_cat_Property	Operativo	Numérica	Nº de incidencias relativas a la propiedad
N_cat_Maintenance	Operativo	Numérica	Nº de incidencias de mantenimiento
N_cat_Noise	Operativo	Numérica	Nº de incidencias por ruido
N_cat_Access	Operativo	Numérica	Nº de incidencias relativas al acceso
N_crit_crit	Operativo	Numérica	Nº de incidencias de criticidad crítica
N_crit_high	Operativo	Numérica	Nº de incidencias de criticidad alta
N_crit_med	Operativo	Numérica	Nº de incidencias de criticidad media
N_satisf_1	Operativo	Numérica	Nº de incidencias con satisfacción valorada en 1
N_satisf_2	Operativo	Numérica	Nº de incidencias con satisfacción valorada en 2
N_satisf_4	Operativo	Numérica	Nº de incidencias con satisfacción valorada en 4
N_satisf_5	Operativo	Numérica	Nº de incidencias con satisfacción valorada en 5
ticket_schema_ratio	Operativo	Numérica	Proporción de incidencias resueltas según el esquema previsto
24x7 Access	Operativo	Categórica	Si el emplazamiento dispone de acceso permanente. Categorías: Sí / No / Desconocido
Tipo de Propiedad	Emplazamiento	Categórica	Naturaleza pública o privada de la propiedad. Categorías: Privado / Público / Desconocido
Tipología Propiedad	Emplazamiento	Categórica	Naturaleza jurídica de la propiedad (titularidad). Categorías: 8 categorías + Desconocido
Tipo de Infraestructura	Emplazamiento	Categórica	Tipo de soporte físico (azotea, torre, caseta, etc.). Categorías: 7 categorías + Desconocido

Variable	Bloque	Tipo	Descripción
Uso del Terreno	Emplazamiento	Categórica	Uso del suelo sobre el que se ubica el emplazamiento. Categorías: Agrícola / Residencial / Industrial / Desconocido
Clasificación del terreno	Emplazamiento	Categórica	Clasificación urbanística del terreno. Categorías: Rústico / Urbano / Desconocido
Periodicidad	Contractual	Categórica	Frecuencia de pago de la renta. Categorías: 6 categorías + Desconocido
Forma de renovación	Contractual	Categórica	Mecanismo de renovación (tácita, expresa, sin prórroga). Categorías: 6 categorías + Desconocido
¿Permite Subarrendamiento?	Contractual	Categórica	Si el contrato permite subarrendar el emplazamiento. Categorías: Sí / No / Desconocido
¿Permite Cesión?	Contractual	Categórica	Si el contrato permite la cesión a terceros. Categorías: Sí / No / Desconocido
Valid duration	Contractual	Categórica	Si el contrato se encuentra dentro de su plazo de vigencia inicial. Categorías: Verdadero / Falso
¿Prórroga?	Contractual	Categórica	Si el evento corresponde a un periodo de prórroga del contrato. Categorías: Verdadero / Falso

Tabla 3.1. Diccionario de variables bloque. Elaboración propia.

La variable objetivo se construye a partir de la variable de estado del evento, que presenta originalmente cinco categorías. Se adopta una formulación binaria que agrupa como continuidad los eventos renovados (primera renovación, renovación automática y renovación de arrendamiento), que suman el 78,6% de los casos, y como pérdida los eventos no renovados o rescindidos, que representan el 21,4% restante. Esta binarización refleja con precisión la decisión de negocio relevante: para el equipo de gestión de cartera, lo único que importa es distinguir los emplazamientos que se conservan de los que se pierden, con independencia del mecanismo concreto de renovación.

3.2. Preprocesamiento de datos

El análisis exploratorio reveló varias incidencias en la calidad de los datos que requirieron tratamiento antes del modelado. En primer lugar, un grupo de variables numéricas se encontraba almacenado como texto, con coma decimal y unidades de medida adheridas al valor

(por ejemplo, la renta anual, la duración media de los tickets o la superficie contratada); estas variables se convirtieron a formato numérico mediante una rutina de limpieza específica.

En segundo lugar, el análisis de valores ausentes identificó un conjunto de variables del bloque del emplazamiento con porcentajes de ausencia superiores al 90% (el número de mástiles, la propiedad de la torre, la altura del edificio y los metros cuadrados libres en punta), que se descartaron por carecer de cobertura suficiente para resultar informativas. Para el resto de las variables, los valores ausentes se imputaron mediante la mediana en el caso de las numéricas y mediante una categoría específica de desconocido en el caso de las categóricas, evitando así descartar observaciones completas y permitiendo que la propia ausencia de un dato pudiera aportar información al modelo.

En tercer lugar, se prestó especial atención a la posible fuga de información, esto es, a la presencia de variables que codifican el desenlace que se pretende predecir. Dos variables resultaron problemáticas en este sentido: el estado administrativo del evento y la fecha de rescisión, que solo existe cuando el contrato efectivamente se ha perdido. Ambas se excluyeron del conjunto de variables explicativas, junto con los identificadores y las fechas en formato texto ya resumidas en variables de antigüedad. Su inclusión habría inflado artificialmente el rendimiento del modelo sin aportar capacidad predictiva real sobre eventos futuros.

Más allá de la exclusión de variables que codifican directamente el desenlace, la naturaleza predictiva del modelo exige considerar la dimensión temporal de la información empleada, esto es, garantizar que las variables explicativas reflejen únicamente información disponible antes de que el propietario tome su decisión. Las variables contractuales y del emplazamiento son estructuralmente estables y no plantean este problema, ya que describen condiciones fijadas en el origen del contrato. Las variables operativas, en cambio, recogen el historial acumulado de incidencias del contrato, por lo que su interpretación predictiva descansa en el supuesto de que dicho historial es anterior al evento de renovación analizado. En el marco de aplicación previsto para SiteGuard, este supuesto se satisface por diseño: el modelo se ejecutaría sobre la cartera en la ventana de meses previa a cada vencimiento, empleando como variables el historial de incidencias acumulado hasta ese momento de corte, anterior por construcción a la decisión. Conviene, no obstante, reconocer explícitamente que el conjunto de datos disponible recoge las incidencias como agregados históricos sin marca temporal individual, lo que impide verificar con plena certeza, para cada evento, que ninguna incidencia se registrara en el intervalo inmediatamente posterior a la decisión. Esta cautela, intrascendente para la validación

metodológica presentada, resulta crítica en la implantación operativa, donde la definición precisa del momento de corte de los datos constituye un requisito de diseño del producto.

Por último, se trataron los valores extremos de las variables numéricas mediante winsorización en el percentil 99. Esta técnica consiste en fijar un techo a los valores que superan dicho percentil, sustituyéndolos por el valor del umbral en lugar de eliminar la observación completa. El caso más ilustrativo es la renta anual, cuya distribución presentaba un máximo de 1.111.111 euros frente a una mediana de apenas 6.817 euros, indicativo de unos pocos valores atípicos capaces de distorsionar los coeficientes de un modelo lineal. La figura 3.1 muestra el efecto de la winsorización sobre esta variable, que comprime el rango sin alterar la masa central de la distribución. Este tratamiento es especialmente pertinente en la regresión logística, sensible a los valores extremos, a diferencia de los modelos basados en árboles, que son robustos a ellos por su propia construcción.

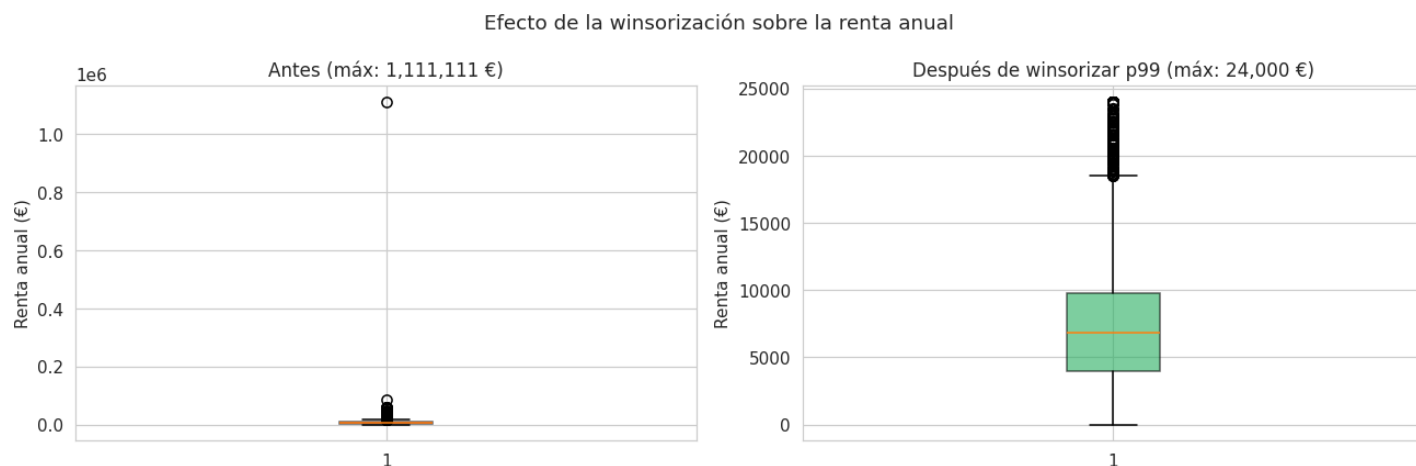


Figura 3.1. Efecto de la winsorización en el percentil 99 sobre la renta anual. Elaboración propia.

3.3. Selección de variables y tratamiento de la multicolinealidad

A partir del conjunto depurado se seleccionaron las variables explicativas atendiendo a su relevancia conceptual y a su disponibilidad. El análisis de correlación detectó dos pares de variables con correlación de Spearman superior a 0,8, indicativa de multicolinealidad: el número total de tickets con el número de tickets cerrados ($r = 0,99$) y la duración media de los tickets con su duración máxima ($r = 0,95$). En cada par se conservó una sola variable (el número total de tickets y la duración media, respectivamente) para evitar la inestabilidad de los coeficientes que la multicolinealidad introduce en los modelos lineales. El conjunto final de variables explicativas comprende veintisiete variables numéricas y doce variables categóricas.

3.4. Estrategia de validación y gestión del desbalanceo

El conjunto de datos se dividió en un subconjunto de entrenamiento (75% de las observaciones) y uno de prueba (25% restante), empleando un muestreo estratificado que preserva la proporción de la clase pérdida en ambos subconjuntos. El subconjunto de prueba se reservó íntegramente para la evaluación final, de modo que ninguna decisión de modelado se tomó observándolo. Adicionalmente, sobre el subconjunto de entrenamiento se aplicó una validación cruzada estratificada de cinco particiones para verificar la estabilidad del rendimiento del modelo y descartar que los resultados dependieran de una partición concreta.

El desbalanceo de la variable objetivo se gestionó mediante la ponderación de clases. En lugar de generar observaciones sintéticas o de submuestrear la clase mayoritaria, se asignó a cada clase un peso inversamente proporcional a su frecuencia, calculado como el cociente entre el número total de observaciones y el producto del número de clases por la frecuencia de cada clase. Con la distribución observada, este criterio asigna un peso aproximado de 0,64 a la clase de continuidad y de 2,33 a la clase de pérdida, de modo que los errores cometidos sobre la clase minoritaria penalizan al modelo en torno a 3,7 veces más que los cometidos sobre la mayoritaria. Esta estrategia, sencilla y transparente, presenta dos ventajas frente a las alternativas de remuestreo: no distorsiona la distribución real de los datos ni introduce riesgo de fuga de información entre entrenamiento y prueba, y refleja además de manera natural la asimetría de costes del problema de negocio, donde dejar de detectar una pérdida es mucho más costoso que revisar un contrato que se habría renovado.

3.5. Modelos y criterios de evaluación

Se entrenaron dos modelos sobre la misma arquitectura de preprocesamiento. El primero, denominado modelo de referencia o baseline, es una regresión logística sin regularización efectiva ni ponderación de clases, que sirve como punto de comparación para cuantificar la aportación de las decisiones metodológicas. El segundo, el modelo final, es una regresión logística con regularización de tipo L2 y ponderación de clases equilibrada. La regularización L2 penaliza la magnitud de los coeficientes, controlando el sobreajuste que el elevado número de variables podría inducir y mejorando la capacidad de generalización del modelo a datos nuevos.

El preprocesamiento se integró en una tubería o pipeline que encadena la imputación de valores ausentes, la estandarización de las variables numéricas (necesaria para que la regularización penalice de forma equitativa a todas las variables con independencia de su

escala) y la codificación one-hot de las variables categóricas, que transforma cada categoría en una variable binaria. La integración del preprocesamiento en la tubería garantiza que todas las transformaciones se ajustan exclusivamente sobre los datos de entrenamiento, evitando cualquier fuga de información desde el conjunto de prueba. La comparación entre ambos modelos se realiza sobre el conjunto de métricas descrito en la sección 2.4, con prioridad analítica sobre la sensibilidad de la clase pérdida y sobre la capacidad de priorización del ranking resultante.

4. Resultados

4.1. Análisis exploratorio y estadística descriptiva

El análisis exploratorio persigue un doble objetivo: comprender la estructura de los datos y orientar las decisiones de modelado. La variable objetivo, en su formulación binaria, presenta el desbalanceo ya señalado, con 9.037 eventos de continuidad y 2.464 de pérdida. La figura 4.1 ilustra tanto la distribución original en cinco categorías como su agregación binaria.

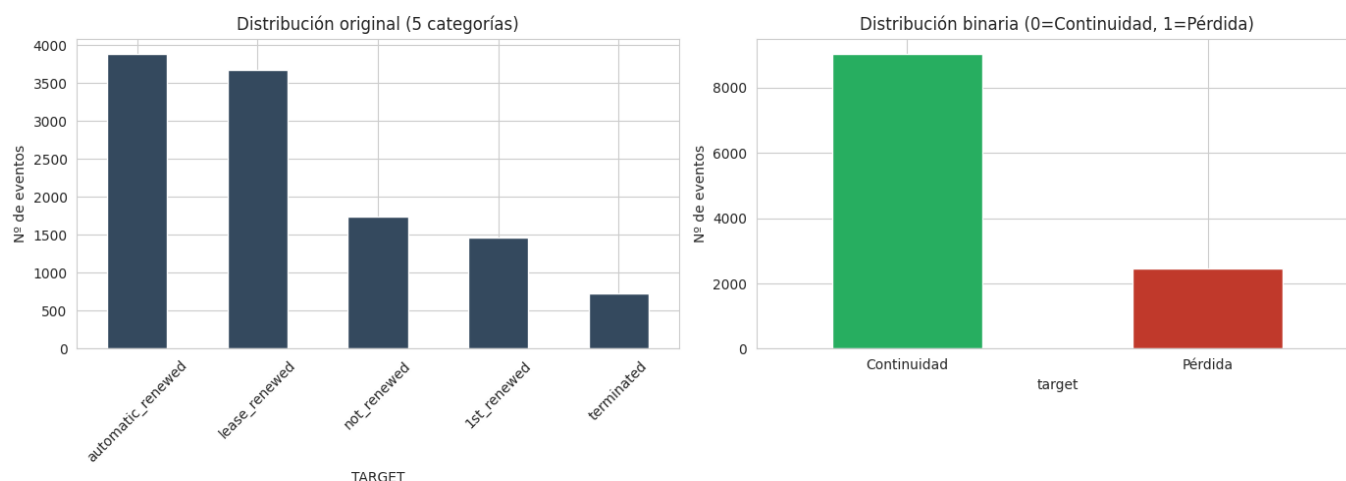


Figura 4.1. Distribución de la variable objetivo. Elaboración propia.

El análisis de valores ausentes, recogido en la figura 4.2, confirmó la presencia de un grupo reducido de variables del bloque del emplazamiento con cobertura insuficiente, descartadas en el preprocesamiento, frente a una mayoría de variables con ausencias moderadas o nulas, especialmente las del bloque operativo, que se encuentran completas.

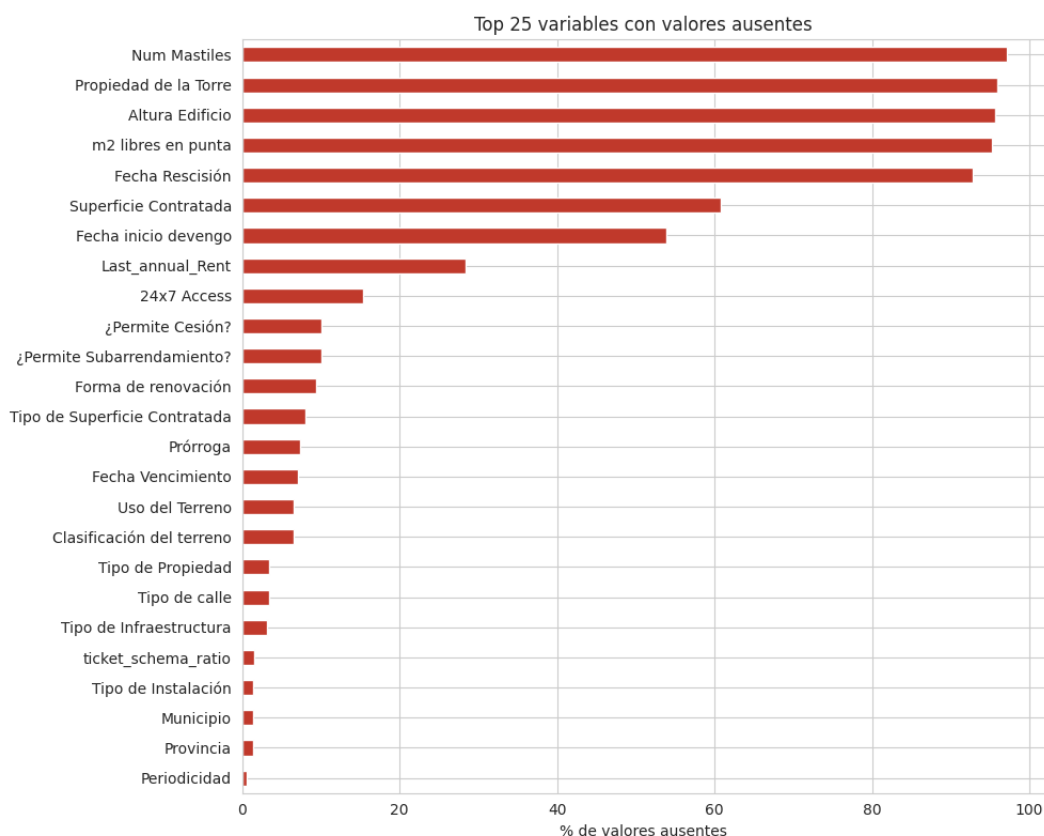


Figura 4.2. Top 25 variables con valores ausentes. Elaboración propia.

El comportamiento de las variables categóricas resulta aún más discriminante. La tabla 4.1 y la figura 4.3 resumen la tasa de pérdida por categoría en las variables más relevantes. Los contrastes son nítidos: los contratos con periodicidad de pago único registran una tasa de pérdida del 66,8%, frente al 15,2% de los de pago anual, y las formas de renovación que no contemplan prórroga tácita multiplican el riesgo respecto a las que sí la contemplan. Estas relaciones anticipan el peso que estas variables tendrán en el modelo.

Variable y categoría	Tasa de pérdida	N eventos
Periodicidad: Pago único	66,8 %	226
Forma de renovación: Sin prórroga	60,9 %	220
Forma de renovación: Expresa por ambas partes	40,0 %	125
Tipología: Land Agg (Telecom Ib.)	32,1 %	187
Periodicidad: Mensual	25,2 %	473
Infraestructura: Marquesina	24,5 %	3.538
Uso del terreno: Agrícola	23,0 %	1.217

Infraestructura: Azotea	19,9 %	7.462
Forma de renovación: Tácita (arrendatario)	8,5 %	272

Tabla 4.1. Tasa de pérdida por categoría en variables seleccionadas. Elaboración propia.

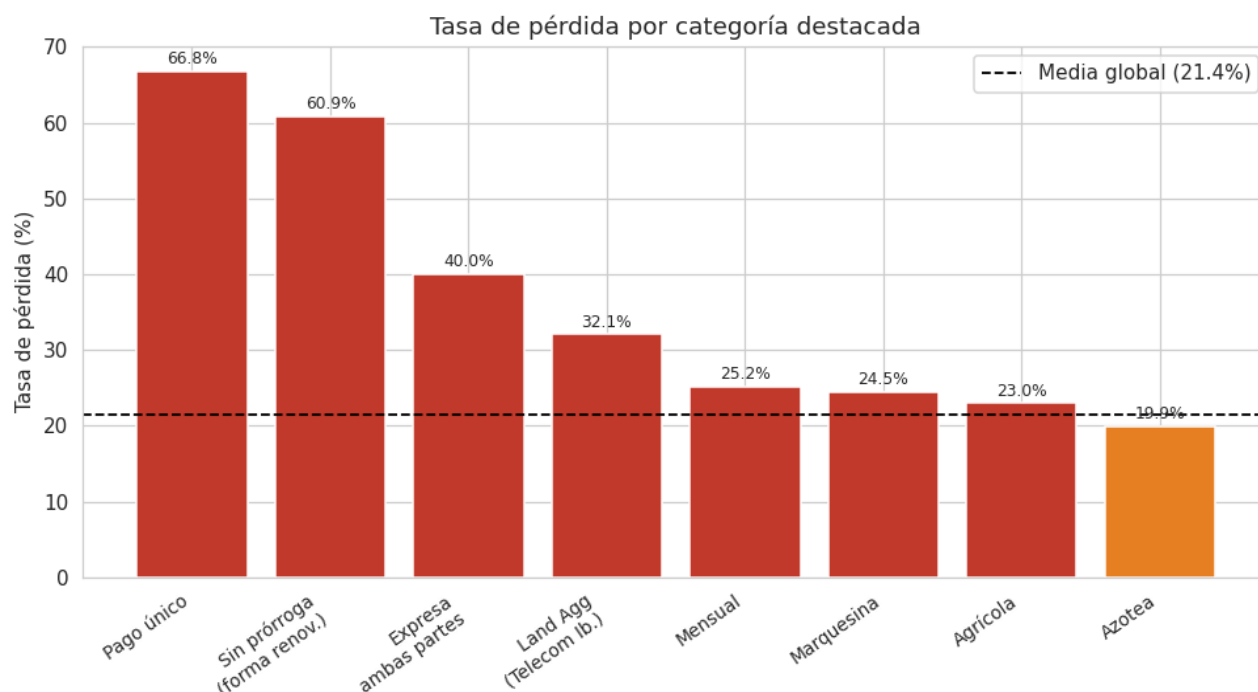


Figura 4.3. Tasa de pérdida por categoría destacada frente a la media global. Elaboración propia.

Para cuantificar la asociación entre las variables numéricas y el desenlace se empleó el coeficiente de correlación de Spearman, adecuado por no asumir relaciones lineales ni distribuciones normales. La figura 4.4 representa la matriz de correlación completa, que cumple una doble función: identificar las variables más asociadas a la pérdida y detectar la multicolinealidad tratada en la metodología.

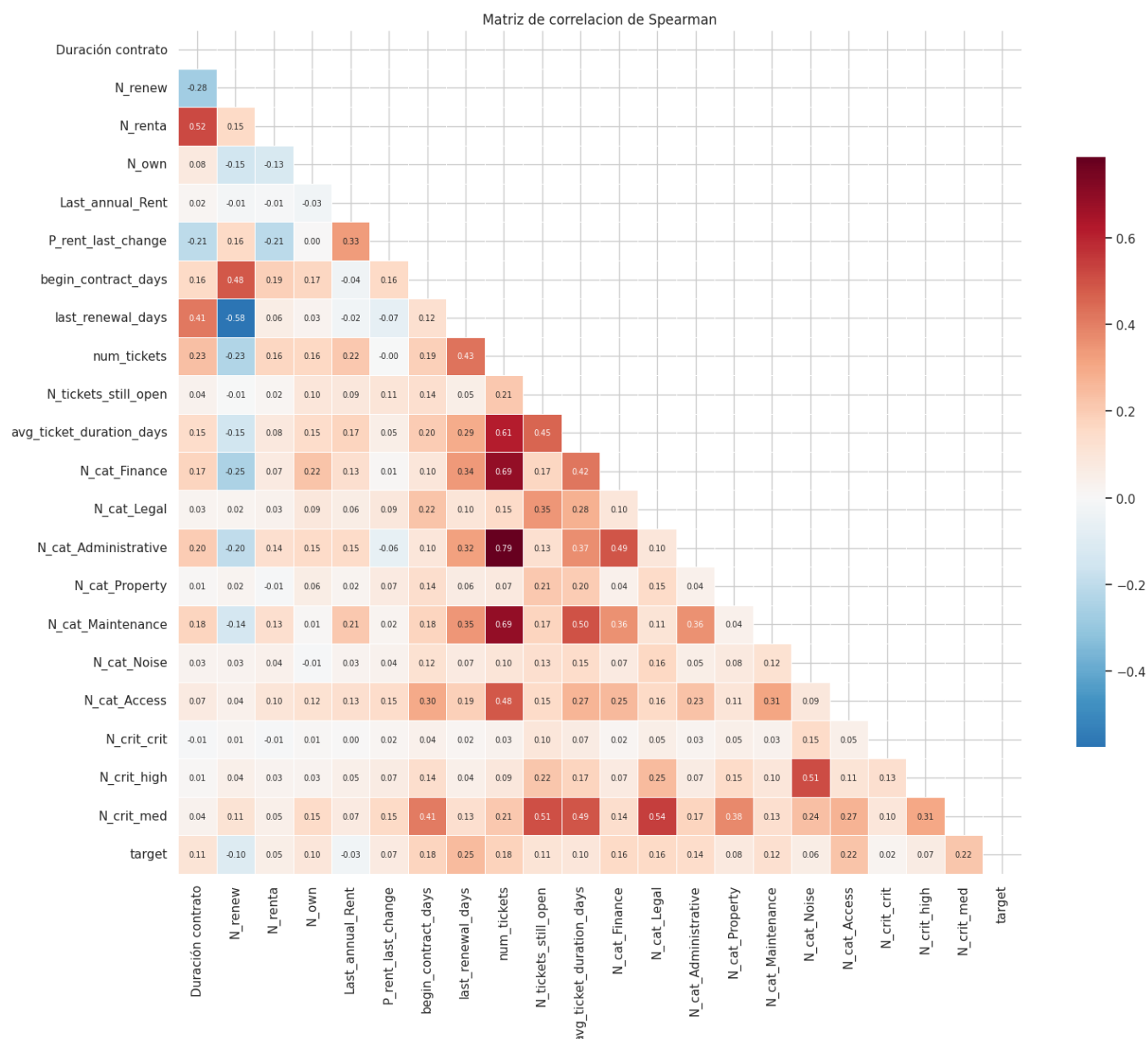


Figura 4.4. Matriz de correlación de Spearman entre variables numéricas. Elaboración propia.

La lectura de la matriz revela que las variables operativas relacionadas con el volumen y la antigüedad de las incidencias muestran las correlaciones positivas más elevadas con la pérdida, aunque de magnitud moderada, lo que es habitual en problemas de clasificación donde ninguna variable individual determina por sí sola el desenlace. En sentido contrario, el número de renovaciones previas presenta una correlación negativa apreciable, indicando que los contratos con un historial largo de renovaciones tienden a mantenerse. La ausencia de correlaciones individuales muy altas con el objetivo refuerza la conveniencia de un modelo multivariante capaz de combinar el efecto conjunto de múltiples señales débiles.

4.2. Comparativa de modelos

La tabla 4.2 compara el rendimiento del modelo de referencia y del modelo final sobre el conjunto de prueba, que no intervino en ninguna fase del entrenamiento. La diferencia entre ambos ilustra el efecto de las decisiones metodológicas adoptadas.

Métrica	Baseline	Modelo final
Exactitud (accuracy)	0,872	0,850
Precisión (clase pérdida)	0,754	0,610
Sensibilidad (clase pérdida)	0,601	0,831
F1 (clase pérdida)	0,669	0,703
AUC-ROC	0,910	0,913
PR-AUC	0,733	0,726
Pseudo-R ² (McFadden)	0,407	0,241

Tabla 4.2. Comparativa de rendimiento sobre el conjunto de prueba. Elaboración propia.

La lectura conjunta de las métricas requiere atender a la asimetría de costes del problema. El modelo de referencia obtiene una exactitud y una precisión ligeramente superiores, pero a costa de una sensibilidad sobre la clase pérdida del 60,1%: deja sin detectar cuatro de cada diez emplazamientos que efectivamente se pierden. El modelo final, gracias a la ponderación de clases, eleva esa sensibilidad hasta el 83,1%, detectando más de ocho de cada diez pérdidas, al precio de un mayor número de falsas alarmas reflejado en la menor precisión. En cuanto al pseudo-R² de McFadden, conviene precisar que esta métrica no constituye un criterio de comparación válido entre ambos modelos: al introducir la ponderación de clases, la función de verosimilitud sobre la que se calcula deja de ser homogénea con la del modelo de referencia, por lo que sus valores no son directamente comparables. Se reporta, por tanto, únicamente como indicador descriptivo del ajuste de cada especificación, y no como base de la decisión, aunque el 0,241 del modelo final sigue siendo un valor aceptable. Los criterios determinantes para la selección del modelo final son la sensibilidad sobre la clase pérdida y la capacidad de discriminación medida por el AUC-ROC, coherentes con el objetivo de negocio de anticipar el mayor número posible de pérdidas.

Para el negocio de SiteGuard, esta es la dirección correcta del intercambio: una falsa alarma cuesta una llamada de gestión innecesaria, mientras que una pérdida no anticipada cuesta entre 55.000 y 165.000 euros. El área bajo la curva ROC, que evalúa la discriminación con independencia del umbral, es prácticamente idéntica en ambos modelos y elevada (0,913),

lo que confirma que la capacidad de ordenación del riesgo no se ve comprometida. La validación cruzada de cinco particiones sobre el conjunto de entrenamiento arrojó un AUC-ROC medio de 0,913 con una desviación típica de 0,004, una variabilidad muy reducida que confirma la estabilidad del modelo.

La figura 4.5 representa las matrices de confusión de ambos modelos. Estas matrices permiten leer directamente el comportamiento en términos de negocio: de cada cuadrante, el de verdaderos positivos recoge las pérdidas anticipadas correctamente (la base del valor de SiteGuard), el de falsos negativos las pérdidas no detectadas (el coste de oportunidad que se busca minimizar) y el de falsos positivos las continuidades señaladas como riesgo (revisiones que suponen únicamente coste de gestión). La figura 4.6 muestra las curvas ROC y de precisión-sensibilidad.

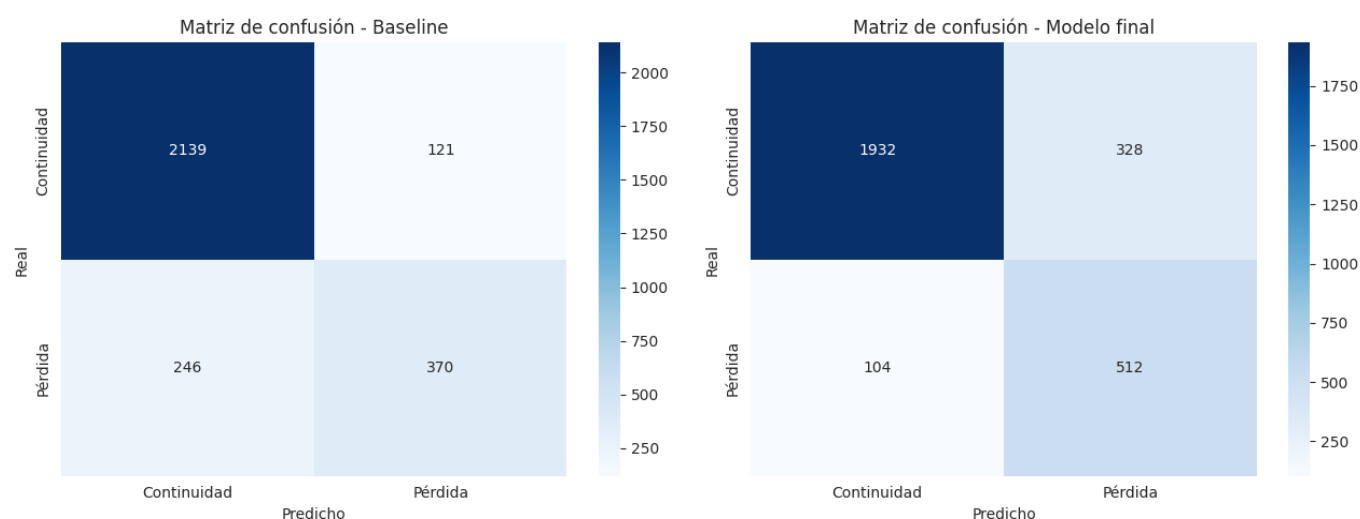


Figura 4.5. Matrices de confusión de ambos modelos sobre el conjunto de prueba. Elaboración propia.

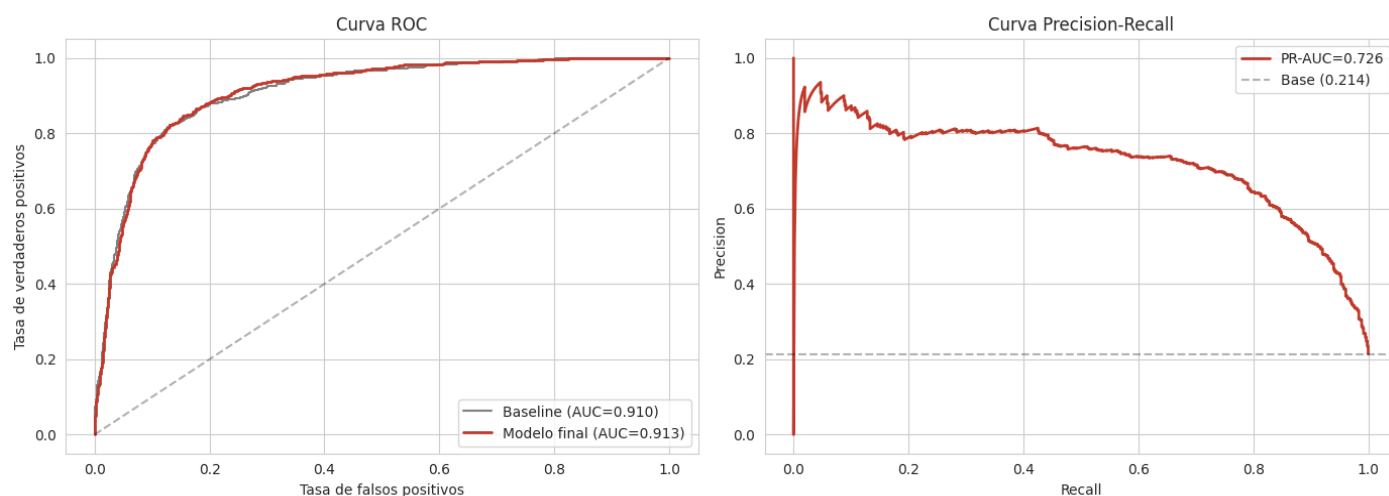


Figura 4.6. Curvas ROC y de precisión-sensibilidad del modelo final. Elaboración propia.

4.3. Interpretación del modelo final

Una de las principales ventajas de la regresión logística es la interpretabilidad directa de sus coeficientes a través de los odds ratios. La tabla 4.3 recoge las variables más influyentes con su coeficiente y su odds ratio asociado, y la figura 4.7 las representa gráficamente.

Variable	Odds ratio	Efecto sobre la pérdida
Forma renov.: Tácita (salvo denuncia arrendatario)	0,012	Reduce drásticamente
N_renew (renovaciones previas)	0,016	Reduce drásticamente
Forma renov.: Tácita (salvo denuncia ambas partes)	0,024	Reduce drásticamente
Forma renov.: Expresa (notificación)	0,109	Reduce
Forma renov.: Expresa (ambas partes)	0,175	Reduce
Periodicidad: Desconocido	3,937	Aumenta
begin_contract_days (antigüedad)	3,683	Aumenta
Tipo infraestructura: Caseta	3,339	Aumenta
Periodicidad: Pago único	3,200	Aumenta
Permite cesión: Sí	0,315	Reduce
Permite subarrendamiento: No	0,326	Reduce
Tipo de propiedad: Público	0,351	Reduce
Tipo de propiedad: Desconocido	0,384	Reduce
Tipo de infraestructura: Indoor	2,359	Aumenta
Valid duration: True	0,454	Reduce

Tabla 4.3. Odds ratios de las variables más influyentes del modelo final. Elaboración propia.

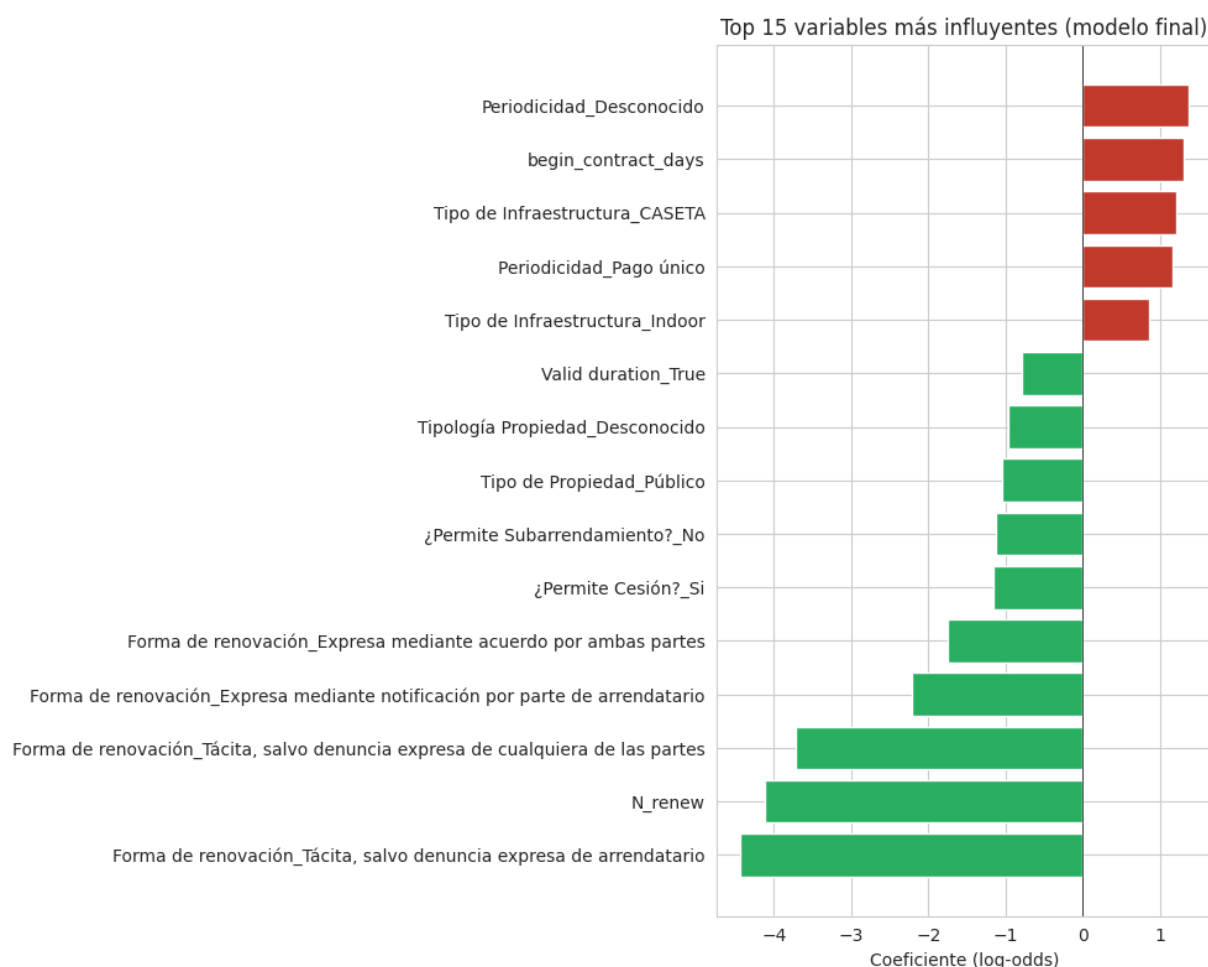


Figura 4.7. Variables más influyentes en el modelo final. Elaboración propia.

Para interpretar correctamente los coeficientes conviene tener presente el preprocesamiento aplicado. Las variables numéricas fueron estandarizadas, de modo que sus coeficientes deben leerse en términos de desviaciones típicas y no de unidades originales: un coeficiente refleja el cambio en el riesgo asociado a un incremento de una desviación típica en la variable, no de una unidad natural. Las variables categóricas, por su parte, se codificaron mediante one-hot encoding, por lo que el odds ratio de cada categoría se interpreta siempre por referencia a la categoría omitida que actúa como base. Por esta razón, la lectura de negocio de los coeficientes debe centrarse en el signo y en la magnitud relativa entre variables, más que en el valor numérico absoluto de cada uno.

Los resultados son coherentes con el análisis exploratorio y con la intuición del sector. Las formas de renovación que contemplan prórroga tácita reducen drásticamente la probabilidad de pérdida, con odds ratios muy inferiores a la unidad: un contrato con renovación tácita salvo denuncia del arrendatario presenta una razón de probabilidades de pérdida cercana al 1% de la de la categoría de referencia, manteniendo el resto de los factores constante. El número de

renovaciones previas opera en la misma dirección protectora, reflejando que cada renovación consolida la relación entre la towerco y el propietario. En el sentido opuesto, la antigüedad del contrato, la periodicidad de pago único y determinadas tipologías de infraestructura incrementan la probabilidad de pérdida, multiplicando la razón de probabilidades por factores superiores a tres. La transparencia de estas relaciones constituye un activo en sí misma: permite al gestor de cartera no solo conocer qué contratos están en riesgo, sino entender por qué, y diseñar en consecuencia la acción correctiva más adecuada.

Conviene asimismo reconciliar la lectura multivariante del modelo con el análisis bivariante previo. Ambos enfoques son coherentes en las señales principales, pero pueden diferir en variables concretas, ya que el análisis bivariante mide la asociación de cada variable con la pérdida de forma aislada, mientras que los coeficientes del modelo reflejan el efecto de cada variable una vez controlado el resto. Una variable que muestra una correlación apreciable de forma aislada puede ver atenuado su coeficiente en el modelo si su información ya está recogida por otras variables correlacionadas con ella; a la inversa, el modelo puede asignar peso a variables cuya relación bivariante con la pérdida era débil pero que aportan información complementaria en presencia del resto. Esta diferencia no constituye una contradicción, sino la consecuencia natural de pasar de un análisis univariante a uno multivariante, y refuerza la conveniencia de un modelo que pondera conjuntamente todas las señales disponibles.

4.4. Aplicación práctica: el ranking de riesgo

Conviene precisar la naturaleza de la salida del modelo. La ponderación de clases empleada para gestionar el desbalanceo, si bien mejora la detección de la clase minoritaria, altera la calibración de las probabilidades estimadas: los valores que el modelo asigna no deben interpretarse como probabilidades absolutas de pérdida en sentido estricto, ya que el reescalado de las clases desplaza sistemáticamente dichas estimaciones. Esta limitación, sin embargo, no afecta a la finalidad del producto. SiteGuard no requiere que la puntuación de un evento represente su probabilidad exacta de pérdida, sino que ordene correctamente los eventos de mayor a menor riesgo, y esa capacidad de ordenación, medida por el AUC-ROC, es independiente de la calibración. Por ello, la salida del modelo se interpreta a lo largo de este trabajo como una puntuación de riesgo orientada a la priorización, y no como una probabilidad calibrada. En caso de que una aplicación futura requiriese probabilidades absolutas fiables, existen técnicas de recalibración posterior, como el escalado de Platt o la regresión isotónica (Platt, 1999; Niculescu-Mizil y Caruana, 2005), que podrían incorporarse sin alterar el modelo subyacente.

El output que SiteGuard entrega al cliente no es una clasificación binaria, sino una ordenación de los eventos de renovación por su puntuación de riesgo, que permite priorizar el esfuerzo del equipo de gestión. Para evaluar la eficacia de esta priorización se ordenaron los eventos del conjunto de prueba por la puntuación de riesgo y se analizó la concentración de pérdidas reales en los tramos de mayor riesgo. La figura 4.8 representa la curva de ganancia acumulada resultante.

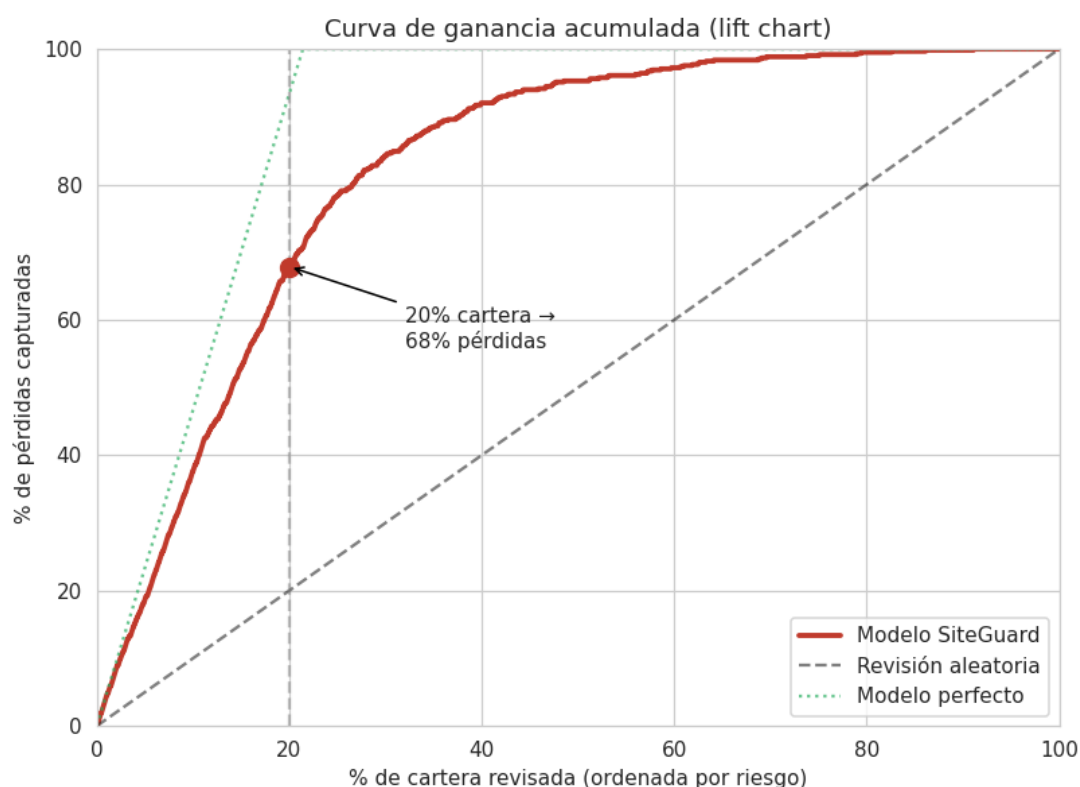


Figura 4.8. Curva de ganancia acumulada (lift chart) del modelo. Elaboración propia.

La eficacia de priorización admite una traducción directa a impacto económico que ilustra el valor de la herramienta. Sobre el conjunto de prueba, compuesto por 2.876 eventos con una tasa de pérdida del 21,4%, cabe esperar en torno a 616 pérdidas. Una revisión del 20% de la cartera seleccionada al azar permitiría abordar unas 123 de esas pérdidas, mientras que la revisión del 20% priorizado por el modelo alcanza 418, lo que supone 294 pérdidas adicionales identificadas a tiempo gracias a la ordenación. Asumiendo que la intervención comercial lograra retener una fracción de esos emplazamientos en riesgo y tomando el rango de coste de pérdida de entre 55.000 y 165.000 euros por emplazamiento establecido en la sección introductoria, el ahorro potencial asociado a esa priorización se situaría, incluso bajo supuestos conservadores de efectividad en la retención, en el orden de varios millones de euros sobre una cartera de este tamaño. Esta magnitud, que debe entenderse como una estimación orientativa y no como una proyección garantizada, dimensiona el retorno económico de la herramienta y

sostiene cuantitativamente la propuesta de valor desarrollada en el Trabajo de Fin de Grado paralelo del proyecto.

Los resultados son contundentes. El 20% de los eventos con mayor probabilidad estimada concentra el 67,7% de las pérdidas reales del conjunto de prueba, lo que equivale a una eficacia de priorización 3,4 veces superior a la de una revisión aleatoria de la cartera. Dicho de otro modo, un equipo de gestión que dedicara su atención únicamente a la quinta parte de la cartera señalada por el modelo abordaría dos de cada tres pérdidas potenciales. La tabla 4.4 detalla la tasa de pérdida real observada en cada decil de riesgo, que decrece de forma monótona desde el 80,6% en el decil de mayor riesgo hasta el 0,3% en el de menor riesgo, evidenciando la capacidad de ordenación del modelo.

Decil de riesgo	Tasa de pérdida real	N eventos
1 (mayor riesgo)	80,6 %	288
2	64,5 %	287
3	35,4 %	288
4	16,7 %	287
5	6,9 %	288
6	4,2 %	287
7	3,5 %	288
8	1,4 %	287
9	0,7 %	288
10 (menor riesgo)	0,3 %	288

Tabla 4.4. Tasa de pérdida real por decil de riesgo. Elaboración propia.

Esta capacidad de priorización es precisamente la propuesta de valor central de SiteGuard: transformar la gestión reactiva de la cartera en una gestión predictiva y focalizada, concentrando los recursos limitados del equipo en los casos de mayor exposición económica.

5. Discusión

5.1. Interpretación de los resultados y conexión con el negocio

Los resultados obtenidos confirman que es posible anticipar con una precisión elevada qué eventos de renovación presentan riesgo de pérdida a partir de las variables que la towerco ya registra en su operativa habitual. El área bajo la curva ROC de 0,913 sitúa el modelo en un rango comparable al de los trabajos de referencia en predicción de abandono en telecomunicaciones (Ahmad et al., 2019), pese a emplear un modelo deliberadamente sencillo e interpretable y a tratar un problema, la decisión de un propietario de suelo, estructuralmente distinto del abandono de un cliente de retail.

Conviene detenerse en el significado de este resultado a la luz de la estrategia metodológica adoptada. La decisión de gestionar el desbalanceo mediante la ponderación de clases, lejos de ser un mero ajuste técnico, resultó determinante para la utilidad del modelo. El modelo de referencia, que no incorporaba esa ponderación, alcanzaba una exactitud agregada superior, pero dejaba sin detectar cuatro de cada diez pérdidas reales, un comportamiento inservible para el objetivo del trabajo. La penalización asimétrica de los errores, que hace que equivocarse sobre una pérdida pese en torno a 3,7 veces más que equivocarse sobre una continuidad, reorientó el modelo hacia la detección de la clase minoritaria y elevó la sensibilidad hasta el 83,1%. Este contraste ilustra una lección metodológica de alcance general: en problemas con clases desbalanceadas y costes asimétricos, la elección de la métrica de optimización y del tratamiento del desbalanceo condiciona la utilidad del modelo tanto o más que la elección del algoritmo.

La interpretación de los coeficientes refuerza la validez del modelo desde una perspectiva de negocio. Las variables más influyentes no son artefactos estadísticos, sino factores con una lógica contractual y operativa reconocible. La forma de renovación emerge como el predictor más potente: las cláusulas de prórroga tácita, en las que el contrato se renueva automáticamente salvo denuncia expresa, reducen drásticamente la probabilidad de pérdida. Este hallazgo es coherente con la naturaleza del problema descrita en la introducción, donde la inercia contractual juega a favor de la towerco: cuando la continuidad es la opción por defecto, hace falta una acción deliberada del propietario para romper el vínculo, y esa fricción protege el emplazamiento. El número de renovaciones previas opera en la misma dirección, confirmando que cada renovación consolida la relación y reduce el riesgo futuro, un patrón consistente con

la idea de que la antigüedad de la relación genera confianza y costes de cambio para ambas partes.

En el sentido opuesto, los factores que incrementan el riesgo dibujan un perfil de contrato vulnerable igualmente interpretable. La periodicidad de pago único, la antigüedad elevada del contrato y determinadas tipologías de infraestructura elevan la probabilidad de pérdida por factores superiores a tres. Resulta particularmente revelador que la ausencia de información en ciertas variables, codificada como categoría propia durante la imputación, presente también capacidad predictiva: los contratos sobre los que la towerco no dispone de datos completos tienden a perderse con mayor frecuencia, lo que sugiere que la calidad del registro documental puede actuar como un proxy del grado de atención y seguimiento que recibe cada emplazamiento.

5.2. Implicaciones para el negocio

La aportación más relevante desde la perspectiva del negocio no reside en la métrica agregada de discriminación, sino en la eficacia de priorización del ranking. La capacidad de concentrar cerca de dos tercios de las pérdidas reales en una quinta parte de la cartera traduce directamente el modelo en una herramienta de asignación de recursos: permite al equipo de gestión multiplicar por más de tres el rendimiento de su esfuerzo comercial frente a una revisión no asistida de la cartera. Esta magnitud es la que sostiene la propuesta de valor de SiteGuard y, por extensión, el modelo de monetización basado en el valor generado que se desarrolla en el Trabajo de Fin de Grado paralelo del proyecto.

La cuantificación económica de esta eficacia resulta inmediata a partir de los parámetros del sector. Dado que el coste de pérdida de un emplazamiento se sitúa entre 55.000 y 165.000 euros, la recuperación de un número reducido de pérdidas evitadas gracias a la priorización del modelo compensa con holgura el coste de la herramienta para una cartera de tamaño medio. El valor del modelo no radica, por tanto, en sustituir el criterio del gestor, sino en dirigir su atención: en una cartera de miles de contratos que en su mayoría se renovarían sin intervención, el modelo identifica el subconjunto reducido donde la actuación comercial concentra su rentabilidad, transformando una gestión reactiva y dispersa en una gestión predictiva y focalizada.

La interpretabilidad del modelo añade una segunda dimensión de valor que conviene subrayar. El hecho de que las variables más influyentes sean coherentes con la lógica del sector refuerza la credibilidad del modelo ante interlocutores técnicamente sofisticados, como suelen

ser los equipos de decisión de las towercos, y permite acompañar cada predicción de una explicación accionable. Frente a la opacidad característica de los modelos de caja negra, un modelo que expone los factores que elevan el riesgo de cada contrato no solo predice, sino que orienta la acción correctiva: si el riesgo se asocia a un historial de incidencias de acceso no resueltas, la intervención adecuada difiere de la que correspondería a un contrato vulnerable por su forma de renovación. Esta transparencia es, en un sector conservador y poco habituado a la automatización de decisiones, un factor de adopción tan relevante como el propio rendimiento predictivo.

5.3. Comparación con la literatura

Los resultados de este trabajo dialogan con la literatura sobre predicción de abandono revisada en la sección 2, aunque con matices que conviene precisar. El rendimiento alcanzado se sitúa en línea con los valores documentados en los estudios de referencia del sector de las telecomunicaciones, donde los modelos de predicción de abandono alcanzan áreas bajo la curva ROC en el entorno del 90% (Ahmad et al., 2019; Chang et al., 2024). No obstante, la comparación debe realizarse con cautela, ya que la mayoría de estos trabajos abordan el abandono de cliente final, un problema en el que el comportamiento del consumidor genera abundantes señales de uso continuas, mientras que el problema aquí estudiado se apoya en señales más escasas y de naturaleza fundamentalmente contractual y operativa.

Esta diferencia confiere un valor particular a los resultados obtenidos. Que un modelo interpretable alcance un rendimiento comparable al de la literatura sobre un problema con menor densidad de señales y sobre una clase minoritaria sugiere que las variables contractuales y operativas registradas por la towerco contienen una información predictiva sustancial, hasta ahora infrautilizada. El trabajo confirma así la hipótesis implícita en la revisión de literatura: la madurez metodológica de la predicción de abandono es trasladable a la retención de contratos de arrendamiento de infraestructura, un dominio que no había recibido atención específica pese a su relevancia económica.

Asimismo, el trabajo se distancia deliberadamente de la tendencia dominante en la literatura reciente hacia modelos de complejidad creciente. Mientras buena parte de los estudios persiguen maximizar el rendimiento predictivo mediante técnicas de ensamble o redes neuronales, este trabajo prioriza la interpretabilidad, en coherencia con el contexto de aplicación. Esta elección no implica desconocer el potencial de los modelos más complejos, sino reconocer que en un entorno de decisión empresarial donde la confianza y la posterior

explicación condicionan la adopción, un modelo transparente con un rendimiento ligeramente inferior puede resultar preferible a uno opaco marginalmente superior. La cuantificación de ese posible margen de mejora se plantea, en todo caso, como línea de trabajo futura.

5.4. Limitaciones y líneas de mejora

El trabajo presenta limitaciones que conviene reconocer explícitamente, tanto por rigor metodológico como por su relevancia para una eventual implantación real.

La primera y más relevante es que el modelo se ha entrenado sobre los datos de una única towerco, lo que limita las garantías sobre su capacidad de generalización a carteras con estructuras contractuales, tipologías de emplazamiento o patrones operativos distintos. Un modelo entrenado sobre una cartera concreta puede haber capturado particularidades de esa empresa que no se reproduzcan en otras, fenómeno conocido como sobreajuste al dominio. La validación con datos de clientes adicionales constituye, por tanto, un paso necesario antes de cualquier despliegue comercial, y es precisamente uno de los argumentos del efecto de aprendizaje cruzado entre clientes que articula la defensa competitiva de SiteGuard: cada nueva towerco incorporada no solo amplía el mercado, sino que mejora la robustez y la capacidad de generalización del modelo.

La segunda limitación atañe a la calidad del dato, que impone un límite estructural. La existencia de variables con cobertura insuficiente, descartadas durante el preprocesamiento, y la necesidad de imputar valores ausentes introducen una incertidumbre que solo puede reducirse con datos de mayor calidad en origen. Aunque el tratamiento aplicado es metodológicamente sólido, ningún procedimiento de imputación recupera información que no se registró, por lo que la mejora de los procesos de captura de datos en la towerco constituye una palanca de mejora del modelo tan importante como cualquier refinamiento algorítmico.

La tercera limitación se refiere a los factores no observables que pueden condicionar la decisión del propietario. Variables como la presión vecinal sobre la presencia de antenas, las motivaciones personales del propietario, los cambios en su situación patrimonial o las ofertas de operadores competidores escapan por completo a las variables estructuradas disponibles. Esta limitación es inherente al enfoque y establece un techo teórico al rendimiento alcanzable: una parte de las pérdidas responde a causas que ningún modelo basado en los datos internos de la towerco podría anticipar. El reconocimiento de este techo es importante para calibrar las expectativas sobre la herramienta y para situar su valor en términos realistas.

La cuarta limitación, de carácter metodológico, se refiere a la naturaleza estática del modelo. El planteamiento actual trata cada evento de renovación a partir de una fotografía de variables agregadas, sin incorporar la dimensión temporal de la evolución de las señales. Un contrato cuyo volumen de incidencias se ha disparado en los meses previos a la renovación presenta, plausiblemente, un perfil de riesgo distinto al de uno con el mismo número de incidencias acumuladas de forma estable a lo largo de los años, y el modelo actual no distingue ambos casos. A esta consideración se añade la cautela ya señalada sobre la calibración de las probabilidades y sobre la definición precisa del momento de corte temporal de los datos, ambas relevantes para una implantación operativa rigurosa.

De estas limitaciones se derivan las principales líneas de mejora futuras. La primera es la exploración de modelos más complejos, como el gradient boosting, para cuantificar empíricamente el margen de mejora en rendimiento predictivo que se obtendría a costa de la interpretabilidad, lo que permitiría tomar una decisión informada sobre el equilibrio entre ambos atributos. La segunda es la incorporación de variables temporales que capturen la evolución de las señales operativas en los meses previos al evento de renovación, superando la limitación estática señalada. La tercera es el desarrollo de un módulo de recomendación de acción que, sobre la base de la puntuación de riesgo y de los factores que la explican, sugiera al gestor la intervención correctiva más adecuada para cada caso, dando un paso desde la predicción hacia la prescripción. Y la cuarta es la validación multicliente, que constituye el requisito previo para la generalización del modelo y, simultáneamente, el motor de su mejora continua.

6. Conclusiones

El sector de las infraestructuras pasivas de telecomunicaciones europeo atraviesa una fase de maduración en la que la conservación de la cartera de contratos de arrendamiento se ha convertido en una palanca de creación de valor tan relevante como la captación de nuevos emplazamientos. En este contexto, la pérdida silenciosa de emplazamientos en los eventos de renovación constituye un problema económicamente significativo que, sin embargo, el sector aborda con herramientas y procesos manifiestamente insuficientes. El presente Trabajo de Fin de Grado se planteó como objetivo desarrollar y validar empíricamente un modelo predictivo capaz de anticipar qué eventos de renovación presentan riesgo de pérdida, generando un instrumento de priorización útil para la gestión de cartera. A lo largo de sus secciones se ha demostrado, sobre un conjunto de datos real de 11.501 eventos cedido por una towerco europea, que ese objetivo se ha alcanzado.

A continuación, se valida el cumplimiento de cada uno de los objetivos específicos planteados en la introducción.

1. Caracterización y depuración del conjunto de datos

Se ha realizado un análisis exhaustivo del conjunto de datos original de 71 variables, depurándolo mediante un proceso sistemático que incluyó la conversión de variables mal tipificadas, el descarte de variables con cobertura insuficiente, el tratamiento de valores extremos por winsorización y la imputación de ausencias. De forma destacada, se identificaron y eliminaron las fuentes de fuga de información, tanto las variables que codifican directamente el desenlace como la dimensión temporal de las señales operativas, garantizando el carácter genuinamente predictivo del modelo. El resultado es un conjunto depurado de veintisiete variables numéricas y doce categóricas con trazabilidad metodológica completa.

2. Identificación de las variables asociadas a la pérdida

El análisis exploratorio permitió caracterizar las señales contractuales y operativas vinculadas a la pérdida de un emplazamiento. El análisis de las tasas de pérdida por categoría y la matriz de correlación revelaron que la forma de renovación, el número de renovaciones previas, la antigüedad del contrato, la periodicidad de pago y el volumen de incidencias operativas son los factores más estrechamente asociados al desenlace, hallazgos coherentes con la lógica del negocio y posteriormente confirmados por el modelo.

3. Entrenamiento y evaluación de un modelo de clasificación binaria interpretable

Se ha entrenado una regresión logística con regularización L2 y ponderación de clases, comparada frente a un modelo de referencia para cuantificar la aportación de las decisiones metodológicas. El modelo final alcanzó un área bajo la curva ROC de 0,913, validada de forma estable mediante validación cruzada de cinco particiones, y una sensibilidad sobre la clase pérdida del 83%, gestionando adecuadamente el desbalanceo del problema. La elección de un modelo interpretable permitió, además, explicar el porqué de cada predicción a través de los odds ratios de las variables, atributo de especial valor en un entorno de decisión empresarial.

4. Traducción de la salida en un ranking de riesgo accionable

Se ha demostrado que la ordenación de los eventos por su puntuación de riesgo concentra cerca del 68% de las pérdidas reales en el 20% de la cartera de mayor riesgo, lo que multiplica por más de tres la eficiencia de una revisión no asistida. Esta capacidad de priorización se ha traducido a impacto económico, situando el ahorro potencial en el orden de varios millones de euros sobre una cartera representativa, lo que dimensiona el retorno de la herramienta y conecta el resultado técnico con su valor de negocio.

El presente trabajo realiza una contribución específica y delimitada. Su aportación principal consiste en trasladar las técnicas de predicción de abandono, ampliamente consolidadas en el ámbito del cliente final de las telecomunicaciones, a un dominio que no había recibido atención específica pese a su relevancia económica: la retención de contratos de arrendamiento de suelo en el sector de las infraestructuras. Sobre un conjunto de datos real, el trabajo demuestra empíricamente que las variables que la towerco ya registra en su operativa contienen información predictiva suficiente para anticipar las pérdidas con una precisión comparable a la de la literatura de referencia, y lo hace mediante un modelo deliberadamente interpretable que no solo predice, sino que explica y prioriza. Frente a la alternativa de un modelo de mayor complejidad y menor transparencia, la elección metodológica adoptada responde a la convicción de que, en un sector conservador donde la confianza condiciona la adopción, la posterior explicación es un atributo tan determinante como el rendimiento.

Las cifras presentadas deben entenderse como resultados de una versión preliminar del modelo, susceptibles de evolucionar con la incorporación de datos de mayor calidad y de clientes adicionales, y el trabajo reconoce explícitamente sus limitaciones, en particular las

derivadas del entrenamiento sobre una única cartera y de la existencia de factores de decisión no observables. No obstante, los resultados obtenidos ofrecen una base sólida para afirmar la viabilidad técnica de la solución. El núcleo predictivo validado en este trabajo constituye el fundamento sobre el que se construye la propuesta empresarial desarrollada en el Trabajo de Fin de Grado paralelo de la titulación de Administración y Dirección de Empresas, de modo que ambos trabajos, el técnico y el comercial, convergen en una misma iniciativa. El paso natural siguiente es la validación del modelo sobre carteras de towercos adicionales, que permitiría contrastar su capacidad de generalización y activar el efecto de aprendizaje cruzado que constituye, a un tiempo, la principal vía de mejora del modelo y el eje de la defensa competitiva del producto.

7. Declaración del uso de la inteligencia artificial

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

Por la presente, yo, Carlos Mijangos Ridruejo, estudiante de E2 + Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Modelo Predictivo de Riesgo de Pérdida de Contratos de Arrendamiento en Infraestructuras de Telecomunicaciones ", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Crítico:** Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
5. **Interpretador de código:** Para realizar análisis de datos preliminares.
6. **Constructor de plantillas:** Para diseñar formatos específicos para secciones del trabajo.
7. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
8. **Generador previo de diagramas de flujo y contenido:** Para esbozar diagramas iniciales.
9. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.

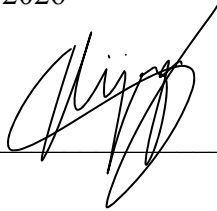
10. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

11. **Traductor:** Para traducir textos de un lenguaje a otro.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 17/06/2026

Firma: _____



8. Bibliografía

- Ahmad, A. K., Jafar, A., y Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. *Journal of Big Data*, 6(28). Obtenido de <https://doi.org/10.1186/s40537-019-0191-6>
- Cellnex Telecom. (2025, 6 de noviembre). *Resultados Cellnex 9M 2025*. Obtenido de https://www.cellnex.com/app/uploads/2025/11/20251106-NDP-Resultados-Cellnex-9M_ES.pdf
- Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., y Benson, V. (2024). Prediction of customer churn behavior in the telecommunication industry using machine learning models. *Algorithms*, 17(6), 231. Obtenido de <https://doi.org/10.3390/a17060231>
- EY-Parthenon. (2024). *The economic contribution of the European wireless infrastructure sector*. European Wireless Infrastructure Association (EWIA). Obtenido de <https://www.cellnex.com/app/uploads/2024/06/ey-parthenon-european-wireless-infrastructure-report-2024.pdf>
- He, H., y Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. Obtenido de <https://doi.org/10.1109/TKDE.2008.239>
- Hosmer, D. W., Lemeshow, S., y Sturdivant, R. X. (2013). *Applied logistic regression* (3.^a ed.). John Wiley & Sons.
- KPR Consult. (2015). *The real costs of decommissioning towers* [Entrevista]. Capacity. Obtenido de <https://capacityglobal.com/news/kpr-consult-the-real-costs-of-decommissioning-towers/>
- LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18), 2395-2399. Obtenido de <https://doi.org/10.1161/CIRCULATIONAHA.106.682658>
- Mijangos, C. (2026). *Plan de Negocio para SiteGuard: Estrategia Comercial de una Solución SaaS Predictiva para Towercos Europeas* [Trabajo de Fin de Grado]. ICADE — Universidad Pontificia Comillas.
- Mordor Intelligence. (2025). *Spain telecom towers market size & share analysis — Growth trends and forecast (2026-2031)*. Obtenido de <https://www.mordorintelligence.com/industry-reports/spain-telecom-towers-market>
- Niculescu-Mizil, A., y Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning (ICML '05)*, 625-632. Obtenido de <https://doi.org/10.1145/1102351.1102430>
- Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. En A. Smola, P. Bartlett, B. Schölkopf y D. Schuurmans (Eds.), *Advances in Large Margin Classifiers* (pp. 61-74). MIT Press.
- S&P Global. (2025, octubre). *Europe towercos slow M&A, eyeing organic expansion and increasing tenancy rates*. Obtenido de <https://www.spglobal.com/market->

[intelligence/en/news-insights/research/2025/10/europe-towercos-slow-manda-eyeing-organic-expansion-and-increasing-tenancy-rates](#)

Steel In The Air. (2025). *Cell tower removal and relocation*. Obtenido de <https://www.steelintheair.com/cell-tower-removal-and-relocation/>

TOTEM. (2025). *TOTEM, la TowerCo industrial en Europa*. Obtenido de <https://www.totemtowers.com/es/>

Zhang, Y. (2025). Research on customer churn prediction model based on machine learning. *Academic Journal of Science and Technology*. Obtenido de <https://drpress.org/ojs/index.php/ajst/article/view/32555>