Facultad de Ciencias Económicas y Empresariales
ICADE

Desarrollo del modelo predictivo de SiteGuard: predicción de la terminación de contratos de arrendamiento en towercos europeas

Autor: Gonzalo Muñoz Navarro
Director: Rafael Castellote Azorín

MADRID | Junio 2026

Resumen

Este Trabajo de Fin de Grado desarrolla una prueba de concepto para predecir la terminación de contratos de arrendamiento de emplazamientos de telecomunicaciones (el suelo sobre el que se asientan torres y antenas). A partir de un conjunto de 11.501 eventos contractuales correspondientes a 4.233 emplazamientos y 4.806 contratos, se construye un pipeline completo de modelización: análisis exploratorio, codificación de variables, un modelo lineal explicativo (regresión logística) y un modelo predictivo basado en gradient boosting (XGBoost). El objetivo es identificar los contratos que el propietario rescinde unilateralmente, un evento poco frecuente (6,3 % de los casos) y de alto impacto económico para la compañía.

Más allá de obtener un modelo con buen poder discriminante, el trabajo pone el énfasis en la honestidad metodológica: se audita el modelo frente a la fuga de información (data leakage) y al sobreajuste, y se demuestra que la validación aleatoria habitual sobrestima el rendimiento en un problema con fuerte componente temporal. Al validar el modelo de forma realista —entrenando con el pasado y evaluando con el futuro— el rendimiento desciende desde un AUC de 0,97 hasta valores en torno a 0,87 de ROC-AUC y 0,53 de PR-AUC, que constituyen la cifra honesta y mantienen al modelo del orden de seis a siete veces por encima de la tasa base de eventos. El trabajo itera el modelo (versiones v1, v2 y v3) hasta una versión depurada, más robusta y defendible, y discute las limitaciones de los datos, en particular la fuerte no estacionariedad de la tasa de terminación y la disponibilidad desigual de la información de incidencias a lo largo del tiempo.

Palabras clave: *churn*, XGBoost, regresión logística, desbalanceo de clases, fuga de información, validación temporal, no estacionariedad, towerco.

Abstract

This Final Degree Project develops a proof of concept to predict the termination of lease agreements for telecommunications sites (the land on which towers and antennas are situated). Using a dataset of 11,501 contractual events corresponding to 4,233 sites and 4,806 contracts, a complete modelling pipeline is constructed: exploratory analysis, variable encoding, an explanatory linear model (logistic regression) and a predictive model based on gradient boosting (XGBoost). The aim is to identify contracts that are unilaterally terminated by the landowner, a rare event (6.3 per cent of cases) with a significant financial impact on the company.

Beyond obtaining a model with good discriminatory power, the work emphasises methodological rigour: the model is audited for data leakage and overfitting, and it is demonstrated that standard random validation overestimates performance in a problem with a strong temporal component. When the model is validated realistically—by training on past data and evaluating against future data—performance drops from an AUC of 0.97 to values of around 0.87 for ROC-AUC and 0.53 for PR-AUC; these figures represent an honest assessment and still keep the model performing six to seven times better than the baseline event rate. The study iterates on the model (versions v1, v2 and v3) until a refined, more robust and defensible version is achieved, and discusses the limitations of the data, in particular the strong non-stationarity of the termination rate and the uneven availability of incident information over time.

Keywords: churn, XGBoost, logistic regression, class imbalance, information leakage, temporal validation, non-stationarity, towerco.

1. Introducción

1.1 Justificación e interés del problema

Los operadores de infraestructuras de telecomunicaciones (towercos) basan buena parte de su negocio en contratos de arrendamiento del suelo donde se ubican sus torres y emplazamientos. Estos contratos son la base de unos ingresos recurrentes y predecibles, de forma análoga a los modelos de suscripción de otros sectores. Por ello, la pérdida de un contrato —su terminación anticipada por parte del propietario del terreno— tiene un impacto directo sobre los ingresos futuros y, además, conlleva costes elevados: relocalización del emplazamiento, pérdida de cobertura, renegociación con el operador móvil y posibles penalizaciones.

En este contexto, anticipar qué contratos están en mayor riesgo de terminación permitiría a la compañía priorizar acciones de retención (renegociación, mejora de condiciones, contacto proactivo con el propietario). El problema es, en esencia, un problema de churn (fuga de clientes), pero con particularidades propias: los contratos son de muy larga duración, el evento de interés es muy poco frecuente y los datos abarcan más de dos décadas, durante las cuales tanto los sistemas de gestión como las condiciones de mercado han cambiado sustancialmente.

1.2 Objetivos

El objetivo general del trabajo es construir y validar de forma rigurosa un modelo capaz de predecir la terminación de contratos de arrendamiento de emplazamientos. Como objetivos específicos se plantean los siguientes:

- Definir adecuadamente la variable objetivo y prevenir la fuga de información derivada de su construcción.
- Construir un modelo explicativo (regresión logística) que permita interpretar el sentido y la magnitud del efecto de cada factor de riesgo.
- Construir un modelo predictivo (XGBoost) y evaluar su capacidad discriminante con métricas adecuadas a un problema desbalanceado.
- Auditar el modelo frente a la fuga de información y el sobreajuste, prestando especial atención a la componente temporal de los datos.
- Iterar y depurar el modelo hasta una versión robusta y defendible, documentando honestamente sus limitaciones.

1.3 Alcance del proyecto

El trabajo es una prueba de concepto centrada en el pipeline de modelización y su validación, no un sistema en producción. No se aborda la integración con los sistemas de la compañía, el despliegue ni el diseño de la interfaz de uso. El alcance se limita a un único conjunto de datos histórico facilitado por un contacto del sector y a la metodología de modelado, validación e interpretación de resultados. Por último, este TFG técnico —correspondiente al Grado en Business Analytics— se desarrolla en paralelo a un TFG de Administración y Dirección de Empresas del mismo autor, centrado en el plan de negocio y la viabilidad comercial de la solución

SiteGuard. Ambos trabajos se conciben como complementarios, de modo que aporten conjuntamente una visión técnica y de negocio integrada del proyecto, sin solapar sus resultados.

1.4 Estructura del documento

Tras esta introducción, el capítulo 2 presenta el marco teórico. El capítulo 3 describe la metodología y los datos. El capítulo 4 recoge los resultados iniciales. El capítulo 5 audita el modelo frente a la fuga de información y el sobreajuste. El capítulo 6 describe la iteración del modelo (versiones v1, v2 y v3). Los capítulos 7 y 8 contienen la discusión y las conclusiones. El documento se cierra con la declaración de uso de IA, la bibliografía y los anexos.

2. Marco teórico

2.1 El churn en negocios de ingresos recurrentes por arrendamiento

El churn o tasa de abandono mide la proporción de clientes (o, en este caso, contratos) que se pierden en un periodo. En negocios de ingresos recurrentes, predecir el churn permite actuar antes de la pérdida. La literatura de churn se ha desarrollado sobre todo en telecomunicaciones móviles y banca (Neslin et al., 2006; Verbeke et al., 2012), pero sus principios son trasladables al arrendamiento de infraestructuras: identificar señales tempranas de riesgo y convertirlas en una probabilidad de pérdida por contrato. La particularidad aquí es la baja frecuencia del evento y la larga vida de los contratos, lo que sitúa el problema a medio camino entre la clasificación binaria y el análisis de supervivencia.

2.2 Modelos predictivos de churn: estado del arte

Los enfoques de churn abarcan desde modelos estadísticos clásicos (regresión logística, análisis de supervivencia) hasta modelos de aprendizaje automático no lineales (random forests, gradient boosting, redes neuronales). En la práctica, los métodos basados en árboles de decisión potenciados (gradient boosting) suelen ofrecer el mejor compromiso entre rendimiento e interpretabilidad para datos tabulares, y son hoy el estándar de facto en competiciones y aplicaciones reales (Chen y Guestrin, 2016). Revisiones recientes del campo confirman el predominio de los métodos basados en ensembles y boosting, pero señalan como reto persistente la limitada capacidad de generalización de los modelos de churn entre dominios y periodos (De y Prabu, 2022), un problema directamente ligado a la no estacionariedad que se aborda en este trabajo. En sectores de ingresos recurrentes análogos al aquí estudiado —como los negocios de alquiler—, los modelos recientes alcanzan valores discriminantes elevados pero acotados (p. ej., un AUC en torno a 0,88 en la predicción de bajas de un negocio de alquiler de electrodomésticos; Suh, 2023).

2.3 Algoritmos de clasificación

La regresión logística modela la probabilidad del evento como una función lineal de las variables a través de la transformación logit; sus coeficientes son directamente interpretables como odds ratios, lo que la hace idónea como modelo explicativo (Hastie, Tibshirani y Friedman, 2009). Los árboles de decisión particionan el espacio de variables de forma recursiva; al combinarse mediante boosting —ajustando cada árbol a los errores del anterior— se obtiene XGBoost, una

implementación eficiente y regularizada de gradient boosting que gestiona de forma nativa los valores ausentes y captura interacciones no lineales (Chen y Guestrin, 2016).

2.4 Desbalanceo de clases: ponderación frente a remuestreo

Cuando la clase de interés es minoritaria, los modelos tienden a ignorarla. Existen dos familias de soluciones: el remuestreo (sobremuestrear la clase minoritaria, p. ej. SMOTE —Chawla et al., 2002—, o submuestrear la mayoritaria) y la ponderación de la función de pérdida (asignar más peso a los errores sobre la clase minoritaria). En este trabajo se opta por la ponderación (class_weight en la regresión logística y scale_pos_weight en XGBoost), evitando SMOTE: el remuestreo sintético puede introducir ejemplos artificiales poco realistas y, como se mostrará, la ponderación no altera la capacidad de ordenación del modelo, solo su punto de operación (He y Garcia, 2009).

2.5 Fuga de información y validación con componente temporal

La fuga de información (data leakage) se produce cuando el modelo accede, durante el entrenamiento, a información que no estaría disponible en el momento real de la predicción, lo que infla artificialmente las métricas (Kaufman et al., 2012). Una fuente habitual y sutil es la validación: en datos con estructura temporal, la validación cruzada aleatoria mezcla pasado y futuro, permitiendo que el modelo "aprenda" patrones de un periodo y se evalúe sobre el mismo periodo. La alternativa correcta es la validación temporal (entrenar con el pasado y evaluar con el futuro), más exigente y representativa del uso real (Bergmeir y Benítez, 2012). Además, cuando un mismo emplazamiento aparece en varios contratos, debe evitarse que filas del mismo emplazamiento caigan a la vez en entrenamiento y test, lo que motiva una partición por grupos.

2.6 Métricas de evaluación en clasificación desbalanceada

La exactitud (accuracy) es engañosa con clases desbalanceadas. El área bajo la curva ROC (ROC-AUC) mide la capacidad de ordenar correctamente positivos y negativos con independencia del umbral. Sin embargo, con eventos muy raros, la curva precisión-exhaustividad (Precision-Recall) y su área (PR-AUC) son más informativas, pues se centran en la clase minoritaria (Davis y Goadrich, 2006; Saito y Rehmsmeier, 2015). El umbral de decisión (por defecto 0,5) puede y debe ajustarse al coste relativo de los errores. En este trabajo se reportan ROC-AUC y PR-AUC como métricas principales, siempre acompañando el PR-AUC de su tasa base de referencia.

2.7 Interpretabilidad y explainability (SHAP)

Los modelos no lineales como XGBoost son cajas negras relativas. Los valores SHAP (SHapley Additive exPlanations) atribuyen a cada variable su contribución a cada predicción de forma consistente con la teoría de juegos cooperativos (Lundberg y Lee, 2017), permitiendo entender tanto la importancia global de las variables como su efecto en casos concretos. Junto con la importancia por ganancia (gain) del propio XGBoost, ofrecen una vía para interpretar el modelo y detectar dependencias sospechosas. En la práctica, la interpretación del modelo en este trabajo se apoyó en la importancia por ganancia (gain) de XGBoost y en el análisis univariante de las variables; el cálculo de valores SHAP se contempló como complemento, pero no llegó a incorporarse en la versión final por razones de alcance, por lo que se deja apuntado como posible extensión del análisis de interpretabilidad.

3. Metodología

3.1 Descripción y origen del dataset

El conjunto de datos procede de un operador de infraestructuras de telecomunicaciones (towerco) y recoge el histórico de eventos contractuales de sus emplazamientos. Tras eliminar un registro mal formado, el dataset contiene 11.501 filas (eventos), correspondientes a 4.233 emplazamientos (sites) y 4.806 contratos de arrendamiento distintos. Cada fila representa un evento en la vida de un contrato (una renovación, una terminación, etc.) y reúne variables del emplazamiento, del contrato y del histórico de incidencias previas a ese evento. El histórico abarca desde 1998 hasta principios de 2026. El conjunto original incluía además algunos identificadores redundantes del emplazamiento (por ejemplo, las columnas "ID Emplazamiento" e "Id Emplazamiento", equivalentes entre sí); se conservó uno de ellos y se descartó el duplicado para evitar redundancia en la preparación de los datos.

3.2 Modelo de datos: estructura jerárquica

Los datos se organizan en tres niveles anidados —emplazamiento (site), contrato (arrendamiento) e incidencias (tickets)— que confluyen en cada evento. La Figura 1 resume esta estructura y las familias de variables asociadas a cada nivel. Un mismo emplazamiento puede tener varios contratos a lo largo del tiempo; sin embargo, cada evento calcula sus variables de forma independiente, por lo que se analizan por separado (con la salvedad, en la validación, de no mezclar el mismo emplazamiento entre entrenamiento y test).

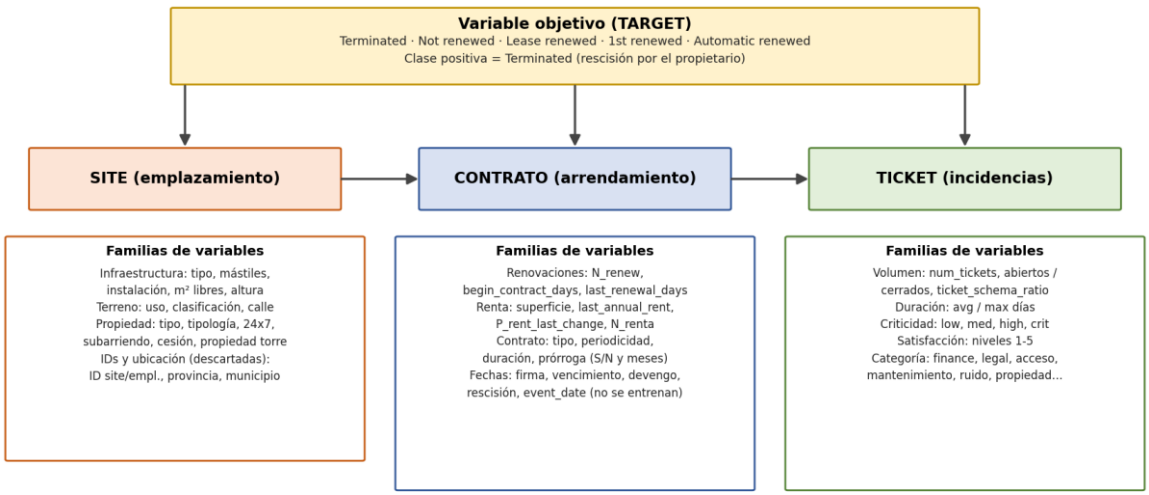

Figura 1. Modelo de datos: estructura jerárquica emplazamiento → contrato → incidencias y familias de variables.

Entre las variables derivadas (de ingeniería de características) destacan varias construidas específicamente para este problema. La Tabla 1 describe las más relevantes. El diccionario de variables (Anexo B) recoge únicamente las variables originales del conjunto de datos; las

variables derivadas empleadas en la modelización —como has_ticket_info, que codifica la disponibilidad de información de incidencias— se documentan en esta tabla y en la descripción metodológica de este capítulo, de modo que todas las variables del modelo quedan descritas en la memoria.

Variable	Descripción
N_renew	Número histórico de renovaciones del contrato. Variable muy informativa.
ticket_schema_ratio	Variable artificial para unificar contratos gestionados por distintos sistemas de ticketing. Vale 0 si las incidencias son solo de 2004-2022 (sin criticidad/satisfacción), 1 si son solo de 2022 en adelante (información completa) y un valor intermedio según la proporción de incidencias de cada periodo.
P_rent_last_change	Último cambio de renta (renta nueva / renta antigua) cercano al evento.
begin_contract_days	Días transcurridos desde el inicio del contrato original.
last_renewal_days	Días transcurridos desde la última renovación.
N_crit_* / N_satisf_*	Recuento de incidencias por nivel de criticidad y de satisfacción (solo disponibles desde 2022).
has_ticket_info	Indicador de si el contrato tiene información de incidencias.

Tabla 1. Principales variables derivadas del conjunto de datos.

3.3 Análisis exploratorio (EDA)

El análisis exploratorio confirmó el fuerte desbalanceo de la variable objetivo (Figura 3) y reveló valores ausentes concentrados en determinadas variables, así como columnas casi vacías (superficie contratada, altura del edificio, metros libres en punta, número de mástiles, propiedad de la torre) que se descartaron. El EDA también puso de manifiesto la fuerte variación temporal de la tasa de terminación, un hallazgo determinante para el resto del trabajo (sección 5.3).

3.4 Limpieza y codificación de variables

Las variables categóricas de texto (tipo de contrato, periodicidad, permisos de subarriendo y cesión, uso y clasificación del terreno, tipo y tipología de propiedad, acceso 24x7, tipo de instalación e infraestructura, tipo de superficie) se codificaron mediante one-hot encoding, tratando el valor ausente como una categoría propia ("Missing"). Las variables numéricas se mantuvieron sin imputar para XGBoost, que gestiona los ausentes de forma nativa; para la regresión logística se imputó la mediana y se estandarizaron. La variable ticket_schema_ratio, ausente en 173 filas sin incidencias, se trató como "sin incidencias" (valor 0 y un indicador has_ticket_info). Por último, en esta fase de limpieza se detectaron y anularon (sustituyéndolos por valores ausentes) dos valores numéricos claramente erróneos —un alquiler anual de 1.111.111 € y un cambio de renta superior a 100 veces—, por tratarse de errores de registro sin sentido económico; su impacto es marginal, pero se corrigen por higiene de los datos (este refinamiento se incorporó en la versión v2 del modelo, sección 6.1).

3.5 Definición de la variable objetivo

La columna TARGET registra el tipo de evento, con cinco categorías: automatic_renewed, lease_renewed, not_renewed, 1st_renewed y terminated. El evento de interés es terminated: la rescisión del contrato por parte del propietario (típicamente comunicada por burofax). Se define así la variable objetivo binaria y = 1 si TARGET = terminated, y = 0 en el resto. Es importante distinguir terminated de not_renewed (no renovación por la otra parte) y de la mera existencia de una fecha de rescisión: un contrato con fecha de rescisión no necesariamente la alcanza, pues puede renovarse antes.

3.6 Prevención de fuga de información

Para evitar fuga, se eliminó del conjunto de variables la propia columna TARGET (que es la fuente de la etiqueta) y la columna Fecha Rescisión. Esta última merece una aclaración: se comprobó que la presencia de fecha de rescisión no implica terminación —aparece en las cinco categorías de TARGET—, por lo que no es la variable objetivo y su inclusión como predictor sería una fuga. Asimismo, los recuentos de incidencias se calcularon utilizando únicamente incidencias anteriores a la fecha del evento. En la misma línea, se descartó también la variable "Forma de renovación": aunque en el análisis exploratorio mostraba una fuerte asociación con la terminación (la categoría "Sin Prórroga" presenta una tasa de terminación muy elevada), describe el modo en que se resuelve el propio evento contractual y queda determinada en el mismo momento del desenlace, por lo que su uso como predictor constituiría una fuga de información —estaría codificando, en gran medida, la propia variable objetivo—. Por ese motivo se excluye del conjunto de predictores, de forma coherente con el criterio aplicado a TARGET y a la fecha de rescisión.

3.7 Estrategia de validación: partición por emplazamiento

Dado que un mismo emplazamiento puede tener varios contratos, una partición aleatoria podría situar contratos del mismo emplazamiento en entrenamiento y test simultáneamente, provocando una fuga sutil. Para evitarlo, la validación cruzada se realizó con GroupKFold (5 particiones) agrupando por identificador de emplazamiento, de modo que ningún emplazamiento aparece a la vez en entrenamiento y test.

3.8 Tratamiento del desbalanceo

Con un 6,3 % de eventos positivos, se aplicó ponderación de clases sin remuestreo: class_weight = 'balanced' en la regresión logística y scale_pos_weight (cociente entre negativos y positivos, ≈ 14,8) en XGBoost. No se utilizó SMOTE.

3.9 Modelos

Se entrenaron dos modelos complementarios. La regresión logística (solver liblinear, ponderación balanceada) cumple un papel explicativo: sus coeficientes, leídos como odds ratios, indican el sentido y la magnitud del efecto de cada variable. XGBoost cumple el papel predictivo, con una configuración deliberadamente regularizada (400 árboles, profundidad máxima 4, tasa de aprendizaje 0,05, submuestreo 0,8 de filas y columnas, min_child_weight 5 y regularización L2).

Estos valores no se fijaron mediante una búsqueda exhaustiva de hiperparámetros, sino en un rango conservador y regularizado habitualmente recomendado para limitar el sobreajuste, adecuado al tamaño del problema (≈ 11.500 registros y más de 50 variables): una profundidad reducida y el submuestreo de filas y columnas acotan la varianza del modelo, mientras que una tasa de aprendizaje baja, compensada con un número moderado de árboles, favorece la estabilidad. Se priorizó así la robustez y la defendibilidad del modelo frente a un ajuste fino que, dada la fuerte componente temporal de los datos, habría exigido una validación temporal específica (se apunta como línea futura en el capítulo 8). La Figura 2 resume el pipeline completo.

Figura 2. Pipeline de modelización: de los datos brutos a la iteración de modelos.

4. Resultados

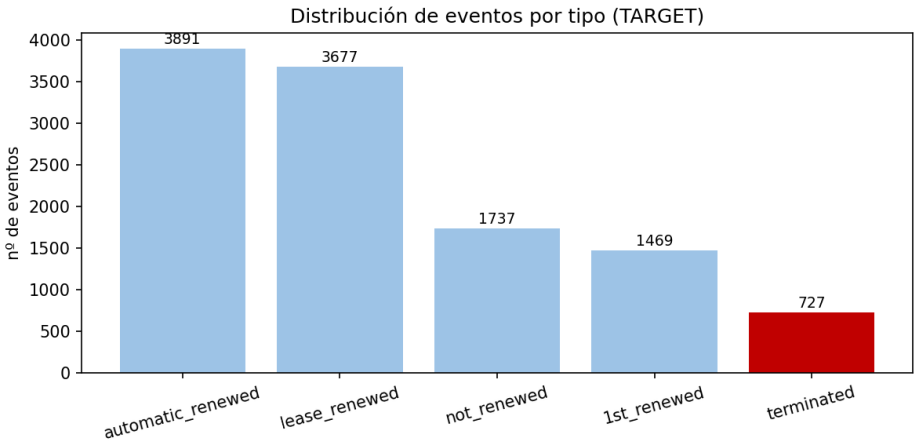

Figura 3. Distribución de eventos por tipo. La clase de interés (terminated) representa solo el 6,3 %.

4.1 Modelo explicativo (regresión logística)

La regresión logística, evaluada con GroupKFold, alcanzó un ROC-AUC de 0,90 y un PR-AUC de 0,48. Sus coeficientes ofrecen una primera lectura de los factores de riesgo. El número de renovaciones históricas (N_renew) destaca como el factor protector más fuerte: los contratos sin renovaciones previas terminan en un 9,2 % de los casos, frente al 3,8 % de los que tienen al menos una, y no se registra ninguna terminación entre los que superan unas quince renovaciones, si bien este último dato procede de un subconjunto muy reducido de contratos y debe interpretarse con cautela. Este patrón monótono es coherente con la idea de que un contrato renovado en múltiples

ocasiones refleja una relación estable y, por tanto, menos propensa a la terminación; como se verá, no constituye fuga de información, sino una señal real.

Más allá de N_renew, la regresión logística señala otros factores con efecto apreciable, interpretables como odds ratios (un valor mayor que 1 aumenta la probabilidad de terminación y uno menor la reduce). La antigüedad del contrato —begin_contract_days— se asocia a un riesgo mayor (odds ratio ≈ 2,7), mientras que la periodicidad de pago único (≈ 0,15) y los emplazamientos de propiedad pública (≈ 0,18) actúan como factores protectores. Ahora bien, varios de los coeficientes más elevados corresponden a indicadores de ausencia de información (por ejemplo, en el tipo de superficie contratada o en el tipo de propiedad), que conviene interpretar como un reflejo de los patrones de disponibilidad de los datos —y no como una relación causal directa—, en la línea de lo que se analiza en el capítulo 7.

4.2 Modelo predictivo (XGBoost) y resumen comparativo

XGBoost mejoró claramente a la regresión logística en la validación con GroupKFold, alcanzando un ROC-AUC de 0,97 y un PR-AUC de 0,82, frente al 0,48 de la logística. La diferencia es esperable: XGBoost captura interacciones no lineales entre las variables que el modelo lineal no puede representar.

4.3 Lectura inicial y necesidad de auditoría

Un ROC-AUC de 0,969 es inusualmente alto para un problema de predicción de abandono: la literatura sobre churn suele reportar valores discriminantes más modestos, habitualmente en el rango 0,75–0,85 (Neslin et al., 2006; Verbeke et al., 2012). Un resultado tan elevado obliga a preguntarse si está inflado por la metodología (fuga de información, sobreajuste o un efecto del balanceo) antes de darlo por bueno. El capítulo 5 somete el modelo a esa auditoría.

5. Auditoría del modelo: ¿hay fuga de información u overfitting?

5.1 ¿Infla el balanceo el rendimiento?

Una primera hipótesis es que el buen resultado se deba a la ponderación de clases. Para descartarlo, se reentrenaron los modelos con y sin balanceo y se compararon las métricas out-of-fold. El ROC-AUC y el PR-AUC apenas variaron (XGBoost: 0,972 con balanceo frente a 0,972 sin él; la logística incluso mejoró ligeramente el PR-AUC sin balanceo). La conclusión es clara: el balanceo no infla las métricas de ordenación —ROC-AUC y PR-AUC son prácticamente insensibles a él—; únicamente reubica el punto de operación, es decir, hace utilizable el umbral por defecto para la clase minoritaria. La capacidad del modelo proviene de las variables, no de la ponderación.

5.2 Overfitting: brecha entre entrenamiento y validación

Para medir el sobreajuste se comparó el rendimiento en entrenamiento frente al de validación en cada partición. La regresión logística prácticamente no sobreajusta (brecha de ROC-AUC de

0,01). XGBoost ajusta el entrenamiento casi a la perfección (ROC-AUC 0,999, PR-AUC 0,976) y muestra una brecha apreciable en PR-AUC (0,98 en entrenamiento frente a 0,82 en validación). Existe, por tanto, cierto sobreajuste; no obstante, el dato que se reporta es el de validación, que es honesto por estar medido sobre emplazamientos no vistos. El sobreajuste no contamina la cifra reportada; solo lo haría si se reportara el valor de entrenamiento.

5.3 Validación temporal: entrenar con el pasado y evaluar con el futuro

La auditoría decisiva es la validación temporal. El diagnóstico previo revela que la tasa de terminación es fuertemente no estacionaria (Figura 4): se mantiene en torno al 3–9 % en la mayoría de años, pero se dispara hasta el 31,8 % en 2022 y desciende después. Además, los años más recientes están censurados por la derecha: las terminaciones de contratos recientes aún no han tenido tiempo de materializarse.

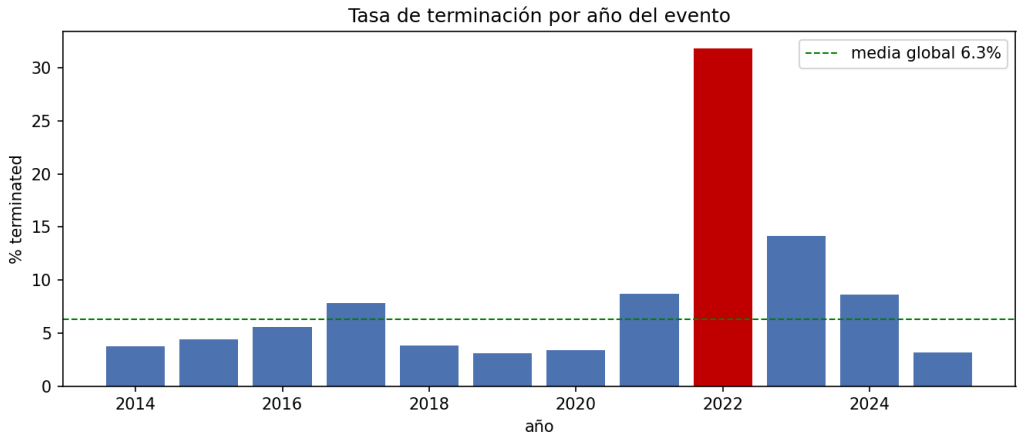

Figura 4. Tasa de terminación por año del evento. Obsérvese el pico anómalo de 2022 y el descenso posterior.

La validación cruzada aleatoria mezcla el año 2022 en entrenamiento y test, de modo que el modelo "aprende" el patrón de ese año y se evalúa sobre él, lo que infla las métricas. Al particionar por fecha del evento (80 % pasado / 20 % futuro, con corte en marzo de 2023) y recalcular el peso de clase solo con el entrenamiento, el rendimiento desciende de forma notable: el XGBoost pasa de un ROC-AUC de 0,86 y un PR-AUC de 0,49 (validación temporal de la versión inicial), frente al 0,97 / 0,82 de la validación aleatoria. La cifra honesta, por tanto, es sustancialmente inferior a la inicial, aunque sigue muy por encima del azar.

5.4 Sondas anti-leakage

Para descartar una fuga grosera se aplicaron dos sondas. En el control negativo se barajó aleatoriamente la variable objetivo del entrenamiento y se reentrenó: el ROC-AUC en test cayó hasta ≈ 0,35, muy lejos del 0,87 real, lo que confirma que la señal del modelo no es un artefacto del pipeline. El valor concreto obtenido (0,35), por debajo del 0,50 teóricamente esperable de un clasificador aleatorio, se explica por la elevada variabilidad del ROC-AUC cuando la clase positiva es muy escasa (6,3 % del conjunto) y se emplea una única permutación del objetivo; lo relevante para descartar la fuga no es que coincida exactamente con 0,50, sino que se aleja por

completo del 0,87 real y se sitúa en el entorno del azar. Repetir el control con varias permutaciones del objetivo permitiría confirmar que el AUC tiende a 0,50 en promedio y acotar su rango de variación. En el análisis univariante se midió la capacidad discriminante de cada variable por sí sola: ninguna variable, ni numérica ni categórica, supera un AUC individual de ≈ 0,68 (la más informativa, N_renew, alcanza 0,63). No existe, por tanto, una única variable "tramposa" que filtre la respuesta; la señal está repartida entre muchas variables débilmente predictivas.

5.5 El factor de confusión de 2022: disponibilidad de información frente a capacidad de predicción

Una preocupación específica es que el modelo "vea mejor" los años recientes por disponer de más información, y no por predecir mejor. Inicialmente se sospechó de la variable ticket_schema_ratio como posible "sello temporal"; sin embargo, los datos no respaldan esa hipótesis: su correlación de Spearman con el año del evento es de apenas 0,03 y el año medio es prácticamente el mismo (≈ 2016) en todos sus tramos. La variable no es un proxy del calendario.

El factor de confusión real está en otra parte. La información de criticidad y satisfacción de las incidencias no existía antes de 2022: el 0 % de los eventos anteriores a 2022 dispone de ella, frente al 38–71 % de los posteriores (Figura 5), con una correlación con el año de 0,56. Esta ruptura se debe a un cambio, en 2022, en el sistema con el que la compañía auditaba los emplazamientos, que pasó a recoger mucha más información sobre el servicio prestado al arrendatario. Como además los contratos con incidencias críticas terminan con más frecuencia (13,1 % frente a 5,1 %), el modelo podría estar asociando "tener información de criticidad" con "terminar", cuando en realidad lo primero es, en gran medida, sinónimo de "ser un evento reciente" del régimen de alta terminación. Estas variables (N_crit_*, N_satisf_* y sus indicadores de ausencia) son, por tanto, las verdaderamente confundidas con el tiempo, y se tratan en el capítulo 6.

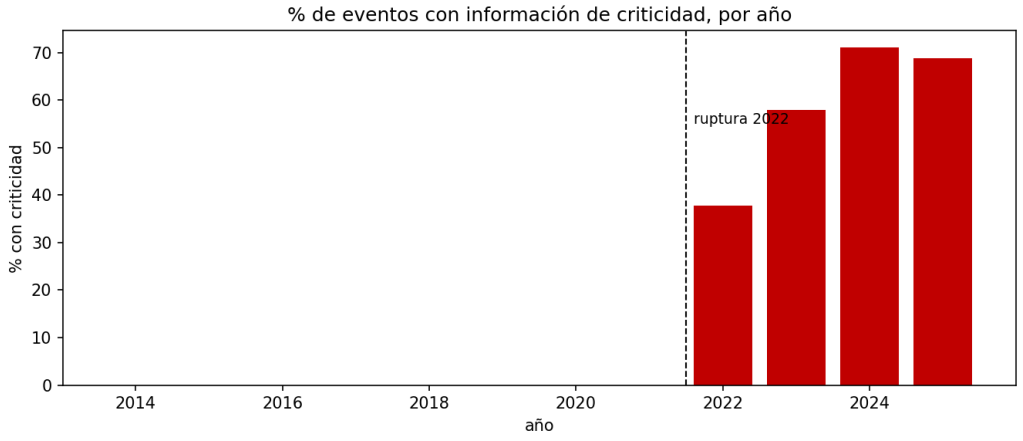

Figura 5. Porcentaje de eventos con información de criticidad por año: ruptura estructural en 2022.

6. Iteración y refinamiento del modelo

6.1 Modelo v1 frente a v2: mejoras aplicadas

A partir de la auditoría se definió una segunda versión del modelo (v2) con tres cambios. Primero, una limpieza de datos puntual: se anularon dos valores claramente erróneos (un alquiler anual de 1.111.111 y un cambio de renta superior a 100 veces), de impacto marginal pero correctos por higiene. Segundo, se completó la auditoría anti-leakage añadiendo el análisis univariante también sobre las variables categóricas. Tercero, y principal, se eliminaron doce variables: el bloque de incidencias confundido con el tiempo (criticidad, satisfacción y sus indicadores de ausencia) y la variable ticket_schema_ratio asociada, que recoge la antigüedad de las incidencias. La Figura 6 resume el conjunto de variables del modelo final y las eliminadas.

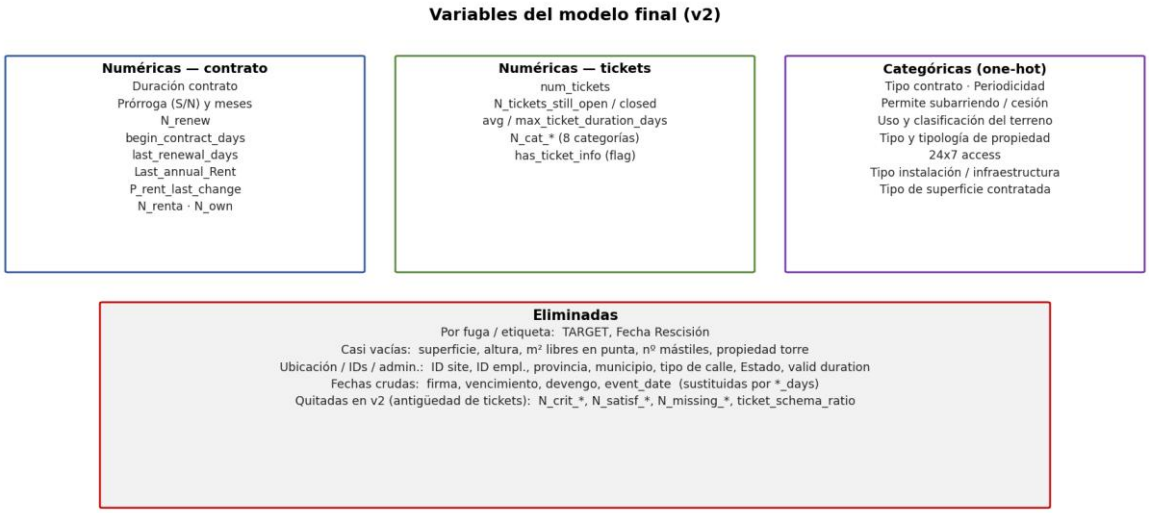

Figura 6. Variables del modelo final (v2) y variables eliminadas, agrupadas por familia.

La comparación entre v1 y v2 (Tabla 2 y Figura 7) muestra un resultado clave: al eliminar las variables sospechosas, el rendimiento en validación aleatoria se mantiene y el rendimiento en validación temporal —el honesto— mejora ligeramente (el PR-AUC del XGBoost sube de 0,49 a 0,55). Es decir, el modelo no dependía del factor de confusión de 2022 para generalizar; al contrario, prescindir de él lo hace algo más robusto. El objetivo de v2 no es un número más alto en la validación aleatoria, sino un modelo más honesto y defendible.

Modelo	ROC-AUC (GroupKFold)	PR-AUC (GroupKFold)	ROC-AUC (temporal)	PR-AUC (temporal)
v1 · XGBoost (completo)	0,972	0,815	0,860	0,487
v2 · XGBoost (depurado)	0,974	0,813	0,871	0,547
v1 · Logística (completo)	0,905	0,479	0,743	0,194
v2 · Logística (depurado)	0,905	0,479	0,762	0,245

Tabla 2. Comparación v1 frente a v2 bajo validación aleatoria (GroupKFold) y temporal.

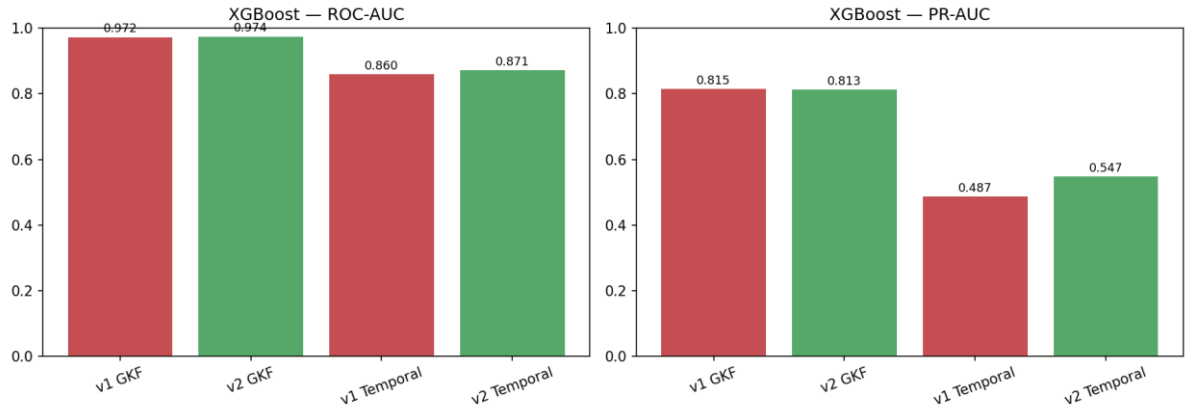

Figura 7. XGBoost v1 frente a v2 en ROC-AUC y PR-AUC, bajo validación aleatoria y temporal.

6.2 Modelo v2 final

El modelo v2 final es un XGBoost entrenado sobre el conjunto depurado de variables. La Figura 8 muestra sus variables más importantes. Resulta llamativo que entre las primeras aparezcan varios indicadores de ausencia ("Missing") de permisos y características del contrato. Esto sugiere que parte de la señal del modelo procede de patrones de datos faltantes, que probablemente reflejan la antigüedad o el origen del contrato más que un factor causal de terminación. No es fuga de información (el análisis univariante lo descarta y la validación temporal lo contabiliza), pero sí una limitación a tener presente en la interpretación.

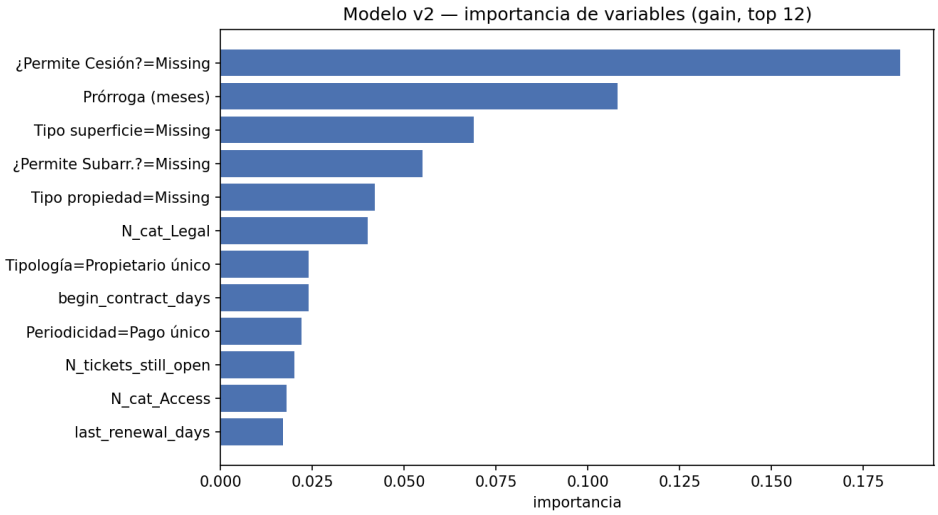

Figura 8. Modelo v2: importancia de variables (ganancia, 12 principales).

6.3 Modelo v3: validación con ventanas temporales

Para evitar leer en exceso un único corte temporal, la versión v3 evalúa el modelo año a año mediante una ventana expansiva: para cada año se entrena con todo el histórico anterior y se predice ese año. La Tabla 3 y la Figura 9 recogen los resultados (se excluye 2025 por su escasez de positivos y la censura). El modelo predice muy bien en los años de régimen estable (ROC-AUC 0,92–0,94) y se debilita en 2022 (ROC-AUC 0,76), precisamente el año del cambio de régimen que el modelo no podía anticipar con datos del pasado. El PR-AUC debe leerse en relación con la tasa base de cada año, que se incluye como referencia.

Año	Nº eventos	% terminación	ROC-AUC	PR-AUC	Mejora sobre azar
2020	715	3,4 %	0,920	0,513	×15,3
2021	392	8,7 %	0,935	0,685	×7,9
2022	497	31,8 %	0,759	0,658	×2,1
2023	729	14,1 %	0,870	0,650	×4,6
2024	998	8,6 %	0,933	0,711	×8,3

Tabla 3. Validación temporal año a año del modelo v2 (ventana expansiva).

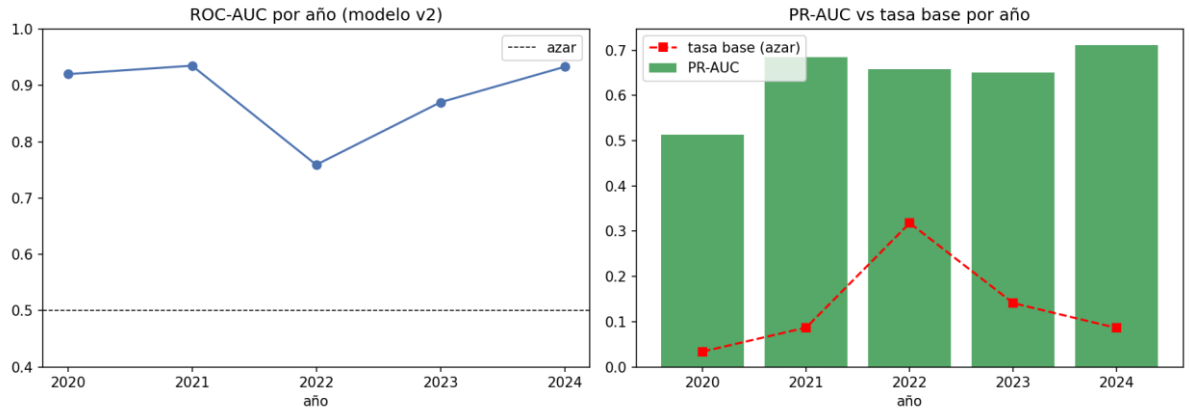

Figura 9. ROC-AUC por año y PR-AUC frente a tasa base, modelo v2 (validación año a año).

7. Discusión

7.1 Qué predice realmente el modelo y para qué sirve

El modelo resultante demuestra un buen desempeño realista, sin indicios de artefactos ni de inflado de métricas, tras la validación rigurosa realizada. La cifra de 0,97 de la validación aleatoria era un espejismo metodológico, no una fuga; la cifra real, validada en el tiempo (ROC-AUC ≈ 0,87, PR-AUC ≈ 0,55, y un rango de 0,76–0,93 según el año), sigue situando al modelo del orden de seis veces por encima del azar, dado que el PR-AUC (≈ 0,55) multiplica por un factor de entre seis y ocho la tasa base de eventos (≈ 6,3 %), en línea con la columna de mejora sobre el azar de la Tabla 3. En términos de negocio, el modelo permite ordenar los contratos por riesgo de terminación y priorizar acciones de retención sobre los más expuestos, aunque a un umbral operativo concreto la precisión es moderada (muchos avisos para capturar la mayoría de las terminaciones), lo que debe tenerse en cuenta al dimensionar el equipo de retención. De forma complementaria, se exploró el ajuste del umbral de decisión —optimizándolo según el F1— para fijar un punto de operación más equilibrado que el umbral por defecto de 0,5, así como una evaluación de la calibración de las probabilidades. Estas comprobaciones refuerzan que, para un uso operativo, conviene seleccionar el umbral en función del coste relativo de los errores y, en su caso, calibrar las probabilidades antes de interpretarlas como riesgo absoluto (véase el Anexo A).

7.2 El efecto de 2022 y las limitaciones de los datos de incidencias

El principal condicionante de los datos es doble. Por un lado, la información de incidencias críticas y de satisfacción solo existe desde 2022, debido a un cambio en el sistema con el que la compañía auditaba los emplazamientos, que pasó a recoger mucha más información sobre el servicio prestado al arrendatario; al estar disponible únicamente en los años recientes, queda confundida con el tiempo, y por eso se excluyó en la versión v2. Por otro lado, la tasa de terminación es fuertemente no estacionaria, con un pico anómalo en 2022 (31,8 %) que no puede explicarse de forma concluyente con los datos disponibles: podría responder a un evento real de negocio (p. ej., las condiciones de mercado de ese año) o, en parte, a un artefacto del mismo cambio de sistema, que pudo mejorar el registro de las terminaciones. Se recomienda investigar este suceso con información adicional de la compañía, dado que esta no estacionariedad afecta

directamente a la capacidad de generalización del modelo. Una limitación adicional es la censura por la derecha: los contratos recientes aún no han tenido tiempo de terminar, lo que sesga a la baja la evaluación de los años más recientes.

7.3 Interpretación de los factores de riesgo

Los factores más estables y de lectura más natural son el historial de renovaciones (a más renovaciones, menos riesgo), la antigüedad del contrato y la presencia de incidencias, especialmente legales. La fuerte importancia de los indicadores de datos faltantes invita a la prudencia: parte de la capacidad predictiva refleja la antigüedad o el origen del contrato más que un mecanismo causal. Para un uso operativo, convendría depurar estas variables de disponibilidad o sustituirlas por información equivalente disponible de forma homogénea en todo el periodo.

8. Conclusiones y líneas futuras

El presente trabajo se planteó un doble objetivo: construir un modelo capaz de predecir la *terminación* —la rescisión del contrato por parte del propietario del emplazamiento, comunicada por burofax— de los contratos de arrendamiento de torres de telecomunicaciones; y, de forma igualmente central, someter ese modelo a una auditoría rigurosa que permitiera discernir si su rendimiento aparente respondía a capacidad predictiva real o a artefactos del propio proceso de modelado. Ambos objetivos se han cumplido. La aportación principal no es, por tanto, una única cifra de rendimiento elevada, sino la distinción —cuidadosamente medida y documentada— entre el resultado optimista que devuelve una validación ingenua y el resultado honesto y defendible que devuelve una validación temporal correctamente diseñada, sobre un conjunto de 11.501 eventos (4.233 emplazamientos y 4.806 contratos) con un 6,3 % de terminaciones.

8.1. Cumplimiento de los objetivos

En el plano del modelado, el objetivo se cubrió con dos algoritmos complementarios. Conviene dejar constancia de una decisión metodológica: aunque el planteamiento inicial mencionaba una «regresión lineal», la naturaleza binaria de la variable objetivo y la necesidad de ponderar las clases hicieron que el modelo lineal correcto fuese la regresión logística —lineal en el *log-odds* y con coeficientes interpretables como *odds ratios*—, reservándose al XGBoost el papel de modelo predictivo de máximo rendimiento. La logística explica; el XGBoost predice. El segundo objetivo, la auditoría, se materializó en el análisis del efecto del balanceo, el diagnóstico de sobreajuste, la validación temporal *out-of-time*, dos sondas anti-fuga (control negativo y AUC univariante) y la depuración de las variables sospechosas de actuar como sello temporal (capítulos 5 y 6). Ello permite responder con evidencia, y no con afirmaciones, a las preguntas que un tribunal puede plantear sobre la solidez del modelo.

8.2. La cifra honesta frente a la cifra optimista

Bajo validación cruzada con *GroupKFold* por emplazamiento, el XGBoost alcanza un ROC-AUC en torno a 0,97 y un PR-AUC de 0,82, frente al 0,90 y 0,48 de la regresión logística. Estas cifras, sin embargo, son optimistas. La validación verdaderamente exigente —y la que debe presentarse como métrica principal en un problema de *churn*— es la temporal, que entrena con los contratos más antiguos y evalúa sobre los más recientes, reproduciendo el escenario real de producción.

Bajo ese esquema, el rendimiento honesto del XGBoost se sitúa en un ROC-AUC de aproximadamente 0,87 y un PR-AUC de entre 0,49 y 0,55 según la versión, como recoge la Tabla 2. La diferencia entre el 0,97 del muestreo aleatorio y el 0,87 temporal no es un detalle, sino el dato central del trabajo.

Para ilustrar qué significa ese rendimiento en términos operativos, en el corte temporal de marzo de 2023 la matriz de confusión a umbral 0,5 detecta 175 de las 189 terminaciones reales (un *recall* del 92 %) a costa de una precisión del 15 %. El modelo es, por tanto, mucho más útil como herramienta de ordenación del riesgo —priorizar qué contratos vigilar— que como clasificador binario a umbral fijo, una matización que condiciona cómo dimensionar el equipo de retención.

8.3. Lecciones metodológicas

- **El balanceo de clases no fabrica señal.** Entrenar con y sin ponderación apenas mueve las métricas de ordenación (en el XGBoost, ROC-AUC de 0,9725 a 0,9717); el balanceo solo reubica el punto de operación, haciendo utilizable el umbral por defecto sobre la clase minoritaria. Reportar ROC-AUC y PR-AUC es, en consecuencia, plenamente legítimo.
- **Existe sobreajuste moderado, pero no contamina la métrica reportada.** El XGBoost ajusta el entrenamiento casi a la perfección, con una brecha apreciable en PR-AUC respecto a la validación; ahora bien, las cifras que se comunican son ya las de validación sobre emplazamientos no vistos, de modo que ese sobreajuste no infla el resultado.
- **La validación aleatoria engaña cuando el fenómeno no es estacionario.** Es la lección más importante del trabajo: el *GroupKFold* controla la fuga por emplazamiento, pero al repartir los años al azar permite que el modelo «vea» el año atípico de 2022 en entrenamiento y sea evaluado también sobre él. La validación temporal, al separar pasado y futuro, devuelve el número real.
- **La depuración de variables (v2) mejora el rendimiento honesto.** Retirar las doce variables disponibles solo desde 2022 no empeora el modelo; al contrario, eleva el PR-AUC temporal del XGBoost de 0,49 a 0,55. El acierto honesto no procede de «tener más datos desde 2022», por lo que se recomienda el modelo depurado (v2) como modelo principal.
- **Las sondas anti-fuga confirman que la señal es real y está repartida.** El control negativo —barajar la variable objetivo— hunde el ROC-AUC de prueba hasta el entorno del azar, muy lejos del 0,87, y ninguna variable supera por sí sola un AUC de ≈ 0,68 (la más informativa, *N_renew*, alcanza 0,63). No hay un único atributo «tramposo»: la capacidad predictiva está distribuida entre muchas variables débilmente informativas.

8.4. Conclusiones sobre los datos y el dominio

La variable objetivo es fuertemente no estacionaria, y este es el hecho empírico que vertebra todo el análisis. La tasa de terminación se mantiene entre el 3 % y el 9 % en los años normales pero se dispara hasta el 31,8 % en 2022. La validación temporal año a año (Tabla 3) lo refleja con nitidez: el modelo predice muy bien en los años de régimen estable (ROC-AUC de 0,76 a 0,93 utilizando únicamente el pasado) y su punto débil es precisamente 2022 (ROC-AUC de 0,759),

porque se le exigió anticipar un cambio de régimen ausente de su histórico. Esto no es un fallo del modelo, sino una limitación honesta que conviene declarar.

El verdadero factor de confusión de 2022 es la disponibilidad de información, no un sello de fecha en el ratio de incidencias. Durante el análisis se corrigió una hipótesis inicial: *ticket_schema_ratio* no es un sello temporal disfrazado (su correlación con el año es ≈ 0,03). El confound real es que la criticidad y la satisfacción del servicio no existían antes de 2022 —del 0 % de los eventos previos al 38-71 % posterior, con una correlación con el año de 0,56—, por un cambio en el software de auditoría de la compañía. Ahora bien, conviene ser crítico con esa explicación: un cambio en la *recogida* de datos justifica que dispongamos de más información, no que se terminen más contratos. El salto del 32 % admite dos lecturas con consecuencias muy distintas —que las nuevas auditorías destaparan problemas reales que derivaron en terminaciones, o que antes de 2022 las terminaciones se registrasen peor, en cuyo caso el pico sería en parte un artefacto de medición de la propia variable objetivo—. Determinar cuál opera queda pendiente de confirmación con el negocio y constituye una de las cautelas centrales del estudio.

Parte de la capacidad predictiva refleja antigüedad u origen del contrato, no causalidad. En el modelo depurado, varias categorías de dato ausente (*Missing*) figuran entre las más importantes: la categoría *Missing* presenta una tasa de terminación del 37,7 % frente al 2,4 % de la mayoritaria, de modo que la ausencia de dato es, en buena medida, un indicio de contrato antiguo. No constituye fuga —ninguna categoría separa por sí sola y la validación temporal ya «paga» este efecto—, pero conviene anticipar la objeción de que la variable más influyente del modelo refleje, sobre todo, que el contrato es viejo. Esta misma lógica explica que se mantuviera fuera del modelo la variable *Forma de renovación*: pese a figurar como predictor fuerte en el diccionario, su inclusión mejoraba la validación aleatoria pero empeoraba la temporal, porque su «fuerza» procede del confound de disponibilidad; su categoría «Sin Prórroga», en cambio, sí es una señal contractual legítima, aunque enmascarada dentro de la masa de *Missing*.

El historial de renovaciones es una señal robusta y honesta. *N_renew* mantiene una relación casi monótona con la terminación —los contratos con cero renovaciones terminan en un 9,2 % de los casos frente al 3,75 % de los que cuentan con alguna, y los que superan las quince no aparecen nunca como terminados—, con solapamiento suficiente (226 contratos terminados con *N_renew* > 0) para descartar separación perfecta o fuga. Es, simplemente, una señal fuerte y plausible desde el negocio: un contrato muy renovado es un contrato estable.

- **8.5. Limitaciones**
- **No estacionariedad y cambio de régimen.** El modelo, entrenado sobre un régimen del 3-9 %, no puede anticipar saltos ausentes del histórico, como el de 2022. El rendimiento honesto está condicionado por la estabilidad del fenómeno.
- **Censura por la derecha.** Los contratos más recientes infrarrepresentan las terminaciones porque aún no han «madurado» (3,1 % y solo 21 positivos en 2025), lo que sesga a la baja la evaluación del periodo final; por ese motivo 2025 se excluyó de la validación temporal defendida.
- **Horizonte de predicción.** Las variables se calculan en la fecha del evento; en esa fecha una terminación puede estar ya «en marcha», de modo que el modelo podría estar *detectando lo inminente* más que *prediciendo con antelación*. Al estar las variables

precalculadas, no es posible evaluar un horizonte mayor sin recomputarlas desde los registros en bruto.

- **Confound de disponibilidad de información.** Parte del acierto refleja la disponibilidad de datos (categorías *Missing*, antigüedad, criticidad solo desde 2022) más que pura predicción; se declara abiertamente y no se oculta.
- **Calibración de las probabilidades.** El uso de *scale_pos_weight* infla las probabilidades estimadas. Para un uso de ordenación del riesgo es irrelevante, pero si se interpretasen como riesgo absoluto requerirían una calibración independiente (Anexo A).
- **Origen del pico de 2022 sin confirmar.** Queda pendiente determinar con el negocio si el aumento corresponde a un evento real o, en parte, a una mejora del registro de las terminaciones.
- **8.6. Líneas futuras**
- **Análisis de supervivencia.** El marco propio del *churn* de contratos es el de tiempo-a-evento (modelos de riesgos tipo Cox o *hazard* discreto), que combinan *cuándo* y *si* se produce la terminación y tratan la censura de forma explícita.
- **Reformulación a horizonte fijo.** Cambiar la pregunta a «¿se cancela en los próximos H meses?» mantiene el enfoque de clasificación, captura el «cuándo» y evita la censura empleando solo contratos con al menos H meses de futuro observable.
- **Validación temporal con ventana móvil como estándar.** Generalizar la evaluación año a año ya ensayada, en lugar de un único corte, para obtener el número honesto acompañado de su rango de variación.
- **Confirmación del cambio de régimen de 2022 con el negocio,** a fin de distinguir un evento real de un posible artefacto de medición en la propia variable objetivo.
- **8.7. Síntesis final**

El modelo desarrollado posee un valor predictivo real y demostrable: supera la tasa base del orden de seis a siete veces en la evaluación temporal, resiste las pruebas de fuga de información y mejora su solidez tras la depuración de variables. Al mismo tiempo, no es «espectacular»: el 0,97 de ROC-AUC de la validación aleatoria era un espejismo derivado de una medición inadecuada —no una fuga—, y la cifra que debe presentarse en resumen y conclusiones es la temporal, en torno a 0,87 de ROC-AUC y 0,53 de PR-AUC, con un rango de 0,76 a 0,93 en la validación año a año. La evolución del trabajo a través de tres versiones —de la optimista (v1) a la depurada (v2) y a la validada en el tiempo (v3)— resume con precisión su aportación: no se ha perseguido un número más alto, sino un número más defendible. El estudio concluye que la terminación de los contratos de arrendamiento de emplazamientos puede predecirse con utilidad práctica como herramienta de priorización del riesgo, y que la validación temporal es imprescindible para no sobrestimar esa capacidad.

9. Declaración del uso de inteligencia artificial

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

Por la presente, yo, Gonzalo Muñoz Navarro, estudiante de E2 + Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "Desarrollo del modelo predictivo de SiteGuard: predicción de la terminación de contratos de arrendamiento en towercos europeas", 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. **Constructor de plantillas**: Para diseñar formatos específicos para secciones del trabajo.
5. **Corrector de estilo literario y de lenguaje**: Para mejorar la calidad lingüística y estilística del texto.
6. **Generador previo de diagramas de flujo y contenido:** Para esbozar diagramas iniciales.
7. **Sintetizador** y divulgador de libros complicados: Para resumir y comprender literatura compleja.
8. **Revisor**: Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
9. **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 25 implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 15/06/2026

Firma: _____

10. Bibliografía

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences, 191, 192–213.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Davis, J., & Goadrich, M. (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning, 233–240.

De, S., & Prabu, P. (2022). Predicting customer churn: A systematic literature review. Journal of Discrete Mathematical Sciences and Cryptography, 25(8), 1965–1985.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning (2nd ed.). Springer.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in Data Mining: Formulation, Detection, and Avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 1–21.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems, 30.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models. Journal of Marketing Research, 43(2), 204–211.

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432.

Suh, Y. (2023). Machine learning based customer churn prediction in home appliance rental business. Journal of Big Data, 10(1), 41.

Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, 218(1), 211–229.

11. Anexos

1.1 Anexo A: Diccionario de variables

Este anexo documenta las 71 columnas originales del conjunto de datos facilitado por la *towerco*, agrupadas por familia. Para cada variable se indica su tipo (y el porcentaje de valores ausentes cuando es relevante), una descripción breve y el tratamiento que recibió en la modelización. Las variables derivadas creadas en el preprocesado (has_ticket_info, begin_contract_days, last_renewal_days, etc.) no se repiten aquí: se describen en la Tabla 1 del capítulo 3. La columna "Tratamiento" refleja las decisiones de la versión final del modelo (v2); las doce variables retiradas en v2 por estar confundidas con el tiempo se señalan de forma expresa.

A.1. Identificadores

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
ID de arrendamiento	texto · 4.806 únicos	Identificador del contrato.	Eliminada (no es predictor).
ID Emplazamiento	texto · 4.233 únicos	Identificador del emplazamiento.	Eliminada como *feature*.
ID Site	texto · 4.233 únicos	Identificador del *site*.	No es *feature*: se usa como **clave de agrupación en el GroupKFold** (anti-fuga).
Id Emplazamiento	texto · 4.232 únicos	Tercer identificador de localización, redundante con los anteriores.	Eliminada (duplicado).

A.2. Variable objetivo y estado del contrato

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
TARGET	5 categorías	Desenlace del evento: *automatic_renewed, lease_renewed,*	**Fuente de la etiqueta** (y = 1 si *terminated*). Eliminada de las variables explicativas.

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
		not_renewed, 1st_renewed, terminated.	
Fecha Rescisión	fecha · 93 % NaN	Fecha de rescisión prevista del contrato.	Eliminada por **fuga**: aparece en las cinco categorías y no implica terminación (§3.6).
Estado	Activo / Rescindido / En Curso	Indica si el contrato sigue vigente.	Eliminada: se entrena con todo el histórico, incluidos los contratos expirados.

A.3. Atributos del contrato y del emplazamiento

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
Tipo Contrato	2 cat. (casi constante)	Arrendamiento vs. *Prepayment* (11.499 frente a 2).	Categórica (*one-hot*); aporta poca información por ser casi constante.
Duración contrato	entero · 47 únicos	Duración del contrato.	Numérica. Incluida.
Valid duration	booleana	Columna sin contenido útil.	Eliminada.
¿Prórroga?	booleana	Si el contrato puede prorrogarse.	Numérica (0/1). Incluida.
Prórroga	float · 7 % NaN	Número de días de prórroga.	Numérica. Incluida.
Forma de renovación	5 cat. · 9 % NaN	Cómo se renueva el contrato (tácita, expresa, sin prórroga…).	**Excluida por fuga**: queda determinada en el mismo momento del desenlace y codifica en gran medida la propia variable objetivo (§3.6).

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
Periodicidad	6 cat. · 1 % NaN	Periodicidad de pago: anual, bimensual, mensual, pago único, semestral, trimestral.	Categórica (*one-hot*, con "Missing"). Incluida.
¿Permite Subarrendamiento?	Sí / No · 10 % NaN	Si el contrato permite subarrendamiento.	Categórica. Incluida.
¿Permite Cesión?	Sí / No · 10 % NaN	Si el contrato permite cesión.	Categórica. Incluida (su categoría *Missing* resulta influyente en v2).
Tipo de Superficie Contratada	7 cat. · 8 % NaN	Tipo de superficie contratada.	Categórica. Incluida.
Uso del Terreno	3 cat. · 7 % NaN	Residencial, agrícola, industrial.	Categórica. Incluida.
Clasificación del terreno	Urbano / Rústico · 7 % NaN	Clasificación del terreno.	Categórica. Incluida.
Tipología Propiedad	8 cat.	Propietario único, comunidad, *land agg*…	Categórica. Incluida.
Tipo de Propiedad	Privado / Público · 3 % NaN	Naturaleza de la propiedad.	Categórica. Incluida (propiedad pública actúa como factor protector).
24x7 Access	Sí / No · 15 % NaN	Si el emplazamiento tiene acceso 24x7.	Categórica. Incluida.
Tipo de Instalación	2 cat. (casi constante)	Existente vs. Nuevo (11.328 frente a 3).	Categórica; casi constante.
Tipo de Infraestructura	7 cat. · 3 % NaN	Azotea, marquesina, local técnico…	Categórica. Incluida.

A.4. Variables geográficas

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
Municipio	texto · 1.629 únicos	Municipio del emplazamiento.	Eliminada (cardinalidad excesiva para codificar).
Provincia	texto · 178 únicos	Provincia del emplazamiento.	Eliminada (alta cardinalidad).
Tipo de calle	texto · 54 únicos	Tipo de vía.	Eliminada (alta cardinalidad).

A.5. Variables de superficie y torre (descartadas)

Columnas con porcentajes de valores ausentes muy altos y sin relevancia estadística; se descartan tras confirmarlo en el EDA.

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
Superficie Contratada	texto · 61 % NaN	Superficie contratada.	Eliminada (mayoritariamente vacía).
Altura Edificio	texto · 96 % NaN	Altura del edificio.	Eliminada (casi siempre vacía).
m2 libres en punta	texto · 95 % NaN	Metros libres en punta.	Eliminada (casi siempre vacía).
Propiedad de la Torre	texto · 96 % NaN	Propiedad de la torre.	Eliminada (casi siempre vacía).
Num Mastiles	float · 97 % NaN	Número de mástiles.	Eliminada (casi siempre vacía).

A.6. Variables de fecha

Las fechas crudas no entran al modelo; se sustituyen por sus equivalentes en días (begin_contract_days, last_renewal_days), descritos en la Tabla 1.

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
event_date	fecha	Fecha en que el evento se hace efectivo.	No es *feature*: ordena el **split temporal** y delimita los recuentos de incidencias previos al evento (anti-fuga).
Fecha Vencimiento	fecha · 7 % NaN	Fecha de vencimiento prevista.	Eliminada (fecha cruda).
Fecha Firma Contrato	fecha	Fecha de firma (idéntica a date_begin_contract).	Eliminada; se usa begin_contract_days.
date_begin_contract	fecha	Inicio del contrato (versión cruda).	Eliminada; resumida en begin_contract_days.
date_last_renewal	fecha	Última renovación (versión cruda).	Eliminada; resumida en last_renewal_days.
Fecha inicio devengo	fecha · 54 % NaN	Inicio de devengo.	Eliminada (mayoritariamente vacía).
end_first_period	fecha	Fin del primer periodo del contrato.	Eliminada (fecha cruda).

A.7. Incidencias (*tickets*) y comportamiento

Los *tickets* son casos abiertos ante incidencias con el emplazamiento o con el propietario. **Todos los recuentos consideran únicamente *tickets* anteriores a event_date** (sin fuga temporal).

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
num_tickets	entero	Número de *tickets* del contrato.	Numérica. Incluida.
ticket_schema_ratio	float [0,1] · 1 % NaN	Variable artificial para unificar contratos gestionados por distintos sistemas de *ticketing* (0 = solo 2004-2022; 1 = solo	Incluida en v1; **retirada en v2** junto al bloque temporal.

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
		desde 2022; intermedio = proporción entre periodos).	
N_cat_Finance … N_cat_Others	8 enteros	Recuento de *tickets* por categoría: *Finance, Legal, Administrative, Property, Maintenance, Noise, Access, Others*.	Numéricas. Incluidas (las legales destacan como señal de riesgo).
N_tickets_still_open	entero	*Tickets* aún abiertos.	Numérica. Incluida.
N_tickets_closed	entero	*Tickets* cerrados.	Numérica. Incluida.
avg_ticket_duration_days	float	Duración media de los *tickets* (días).	Numérica. Incluida.
max_ticket_duration_days	entero	Duración máxima de un *ticket* (días).	Numérica. Incluida.
N_crit_crit / high / med / low	4 enteros	Recuento de *tickets* por criticidad. Solo existe desde 2022.	Incluidas en v1; **retiradas en v2** (confundidas con el tiempo).
N_missing_crit	entero	*Tickets* sin información de criticidad.	Incluida en v1; **retirada en v2**.
N_satisf_1 … N_satisf_5	5 enteros	Recuento de *tickets* por nivel de satisfacción (*one-hot* de las reseñas). Solo desde 2022.	Incluidas en v1; **retiradas en v2** (confundidas con el tiempo).
N_missing_satisf	entero	*Tickets* sin información de satisfacción.	Incluida en v1; **retirada en v2**.
N_renew	entero	Número histórico de renovaciones del contrato.	Numérica. Incluida. **Señal protectora robusta** (a más

Variable	Tipo / valores	Descripción	Tratamiento en el modelo
			renovaciones, menos riesgo).
N_renta	entero	Número de cambios de renta del contrato.	Numérica. Incluida.
N_own	entero (1-7)	Número de propietarios del terreno.	Numérica. Incluida.
Last_annual_Rent	float · 28 % NaN	Última renta anual del contrato.	Numérica. Incluida; el valor 1.111.111 € se trató como ausente por ser un error de registro (v2).
P_rent_last_change	float	Cambio de renta más reciente (renta nueva / renta antigua).	Numérica. Incluida; los valores > 100 se anularon por error de registro (v2). Es la única variable que relaciona filas, pero **no es fuga**.