UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MASTER EN FORMACIÓN PERMANENTE EN INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE MÁSTER

Segmentación Data-Driven para *Cross-sell* del Producto Casa Protegida en Nacional Seguros

SHOGO AMO

Dirigido por

Dr. Alberto Partida

CURSO 2025-2026

TÍTULO: Segmentación Data-Driven para *Cross-sell* del Producto Casa Protegida en Nacional Seguros

AUTOR: Shogo Amo

TITULACIÓN: Máster Universitario en Análisis de Datos Masivos (Big Data)

DIRECTOR/ES DEL PROYECTO: Dr. ALBERTO PARTIDA RODRIGUEZ

FECHA: Abril de 2026

RESUMEN

El trabajo fue realizado en conjunto con Nacional Seguros, empresa del sector asegurador boliviano, con el objetivo de mejorar la estrategia de *cross-sell* del producto de seguro de hogar Casa Protegida.
A partir de fuentes transaccionales internas, se construyó un *dataset* analítico de 1,268,176 registros y se definió una variable objetivo tipo binaria que identifica la contratación del producto en una ventana de seis meses. El desbalance resultante de las clases fue extremo, con apenas un 0.12% de casos positivos en el conjunto de evaluación operativo.
Se utilizó la metodología CRISP-DM para entrenar un modelo de propensión con XGBoost, el cual se evaluó a través de métricas apropiadas para contextos desbalanceados: PR-AUC, ROC-AUC y Lift@K. El modelo optimizado obtuvo un ROC-AUC de 0.7701 y un PR-AUC de 0.0044 en el test de corte único —aproximadamente 3.7 veces el valor esperado bajo un clasificador aleatorio (0.0012), siendo estas métricas globales consistentes con el extremo desbalance de clases. El análisis Top-K demuestra su valor operativo: el 5% de clientes con mejor score agrupa una tasa de conversión 4.10 veces superior a la media y captura el 20.5% de los casos positivos observados.
Como resultado se plantea una herramienta de priorización que transforma la gestión de *cross-sell* de un enfoque masivo a uno orientado por datos, con potencial de ser incorporada en los procesos operativos de la compañía.

Palabras clave: *cross-sell*, XGBoost, modelo de propensión, desbalance de clases, priorización comercial, seguros

ABSTRACT

The work was carried out in collaboration with Nacional Seguros, a company in the Bolivian insurance sector, with the aim of improving the cross-sell strategy for the home insurance product Casa Protegida.
Using internal transactional sources, an analytical dataset of 1,268,176 records was constructed, and a binary target variable was defined to identify the purchase of the product within a six-month window. The resulting class imbalance was extreme, with only 0.12% of positive cases in the operational evaluation set.
The CRISP-DM methodology was used to train a propensity model with XGBoost, which was evaluated using appropriate metrics for imbalanced contexts: PR-AUC, ROC-AUC, and Lift@K. The optimized model achieved a ROC-AUC of 0.7701 and a PR-AUC of 0.0044 on the single-cutoff test set—approximately 3.7 times the expected value under a random classifier (0.0012), with these global metrics being consistent with the extreme class imbalance. The Top-K analysis demonstrates its operational value: the top 5% of customers with the best scores group a conversion rate 4.10 times higher than the average and capture 20.5% of the observed positive cases.
As a result, a prioritization tool is proposed that transforms cross-sell management from a mass approach to a data-driven one, with the potential to be incorporated into the company's operational processes.

Key words: cross-sell, XGBoost, propensity model, class imbalance, commercial prioritization, insurance

AGRADECIMIENTOS

A mi esposa y mi hijo, pirales que sostuvieron esta travesía. Ellos fueron la energía, la inspiración, el fuego que me impulsó en cada paso del camino. Con mucho cariño les dedico esta memoria.

A mis padres por la formación, los valores y el apoyo permanente que permitieron que llegara a este punto del camino.

A mi familia, por el cariño y el respaldo constante que acompañaron cada etapa de este camino.

Al Dr. José Luis Camacho Miserendino, presidente del Grupo Empresarial de Inversiones Nacional Vida, por depositar su confianza al otorgarme un desafío significativo para la organización y por su firme apuesta por la transformación basada en datos del grupo.

A Nacional Seguros, por darme el espacio y recursos necesarios para llevar adelante este trabajo como un aporte real al negocio.

A mi tutor, el Dr. Alberto Partida Rodríguez, por la orientación, las observaciones críticas y el acompañamiento técnico durante el desarrollo del Trabajo Fin de Máster.

A la Universidad Europea de Madrid y al equipo docente del Máster en Big Data, por los conocimientos y herramientas que permitieron afrontar con rigor metodológico un reto de esta magnitud.

A mis compañeros de Máster, por el trabajo conjunto a lo largo del programa, por los intercambios de ideas, los feedbacks oportunos y los debates que enriquecieron la forma de enfocar cada reto.

Ad astra per aspera.

TABLA RESUMEN

	DATOS
Nombre y apellidos:	Shogo Tarifa Amo
Título del proyecto:	Segmentación Data-Driven para *Cross-sell* del Producto Casa Protegida en Nacional Seguros
Directores del proyecto:	Dr. Alberto Partida
El proyecto se ha realizado en colaboración de una empresa o a petición de una empresa:	SI
El proyecto ha implementado un producto: (esta entrada se puede marcar junto a la siguiente)	NO
El proyecto ha consistido en el desarrollo de una investigación o innovación: (esta entrada se puede marcar junto a la anterior)	SI
Objetivo general del proyecto:	El objetivo general del proyecto es utilizar un enfoque data-driven para analizar el histórico de ventas, variables demográficas y el comportamiento de compra de los contratantes naturales de Nacional Seguros. Con ello, se busca priorizar clientes con mayor potencial mediante un modelo de propensión, orientando campañas direccionadas de *cross-sell* del producto Casa Protegida.

Índice

RESUMEN 3
ABSTRACT 4
TABLA RESUMEN 6
Capítulo 1. RESUMEN DEL PROYECTO 12
1.1 Contexto y justificación 12
1.2 Planteamiento del problema 12
1.3 Objetivos del proyecto 12
1.4 Resultados obtenidos 12
1.5 Estructura de la memoria 12
1.6 Alineación con los Objetivos de Desarrollo Sostenible 13
Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE 14
2.1 Estado del arte 14
2.2 Contexto y justificación 16
2.3 Planteamiento del problema 17
Capítulo 3. OBJETIVOS 18
3.1 Objetivos generales 18
3.2 Objetivos específicos 18
3.3 Beneficios del proyecto 19
Capítulo 4. DESARROLLO DEL PROYECTO 20
4.1 Planificación del proyecto 20
4.2 Descripción de la solución, metodologías y herramientas empleadas 21
4.3 Recursos requeridos 30
4.4 Presupuesto 31
4.5 Viabilidad 32
4.6 Resultados del proyecto 32
Capítulo 5. DISCUSIÓN 49
5.1 Interpretación y discusión de los resultados 49

5.2 Limitaciones del estudio y adaptaciones metodológicas 50
Capítulo 6. CONCLUSIONES 51
6.1 Conclusiones del trabajo 51
6.2 Conclusiones personales 52
Capítulo 7. FUTURAS LÍNEAS DE TRABAJO 54
Capítulo 8. REFERENCIAS 55
Capítulo 9. ANEXOS 57
9.1 Anexo A. Diccionario de datos del dataset gold 57
9.2 Anexo B. Variables seleccionadas para el modelo 57
9.3 Anexo C. Configuración de GridSearchCV 58
9.4 Anexo D. Métricas completas del modelo 59
9.5 Anexo E. Estructura del entregable comercial 59

Índice de Figuras

Ilustración 1: Cronograma del proyecto. Fuente: Elaboración propia 20

Ilustración 2:Pipeline de extracción y consolidación del Dataset Gold. Fuente: elaboración propia. 24

Ilustración 3:Pipeline analítico de modelado y entregables. Fuente: elaboración propia. 25

Ilustración 4: Curvas ROC y Precision-Recall del modelo base XGBoost, por horizonte temporal (3 y 6 meses) y conjunto de prueba (test completo y test de corte único). Fuente: elaboración propia. 35

Ilustración 5: Top 20 variables por importancia (ganancia) del modelo base XGBoost — horizonte de 6 meses. Fuente: elaboración propia. 36

Ilustración 6: Curvas ROC y Precision-Recall del modelo final optimizado, por horizonte temporal (3 y 6 meses) y conjunto de prueba (test completo y test de corte único). Fuente: elaboración propia. 38

Ilustración 7: Top 20 variables por importancia (ganancia) del modelo final optimizado XGBoost — horizonte de 6 meses. Fuente: elaboración propia. 39

Ilustración 8: Matriz de confusión — Modelo optimizado 6 meses, test de corte único (Top 10%). Fuente: elaboración propia. 41

Ilustración 9: Vista de rendimiento del modelo — Dashboard interactivo en Tableau. Fuente: elaboración propia. 43

Ilustración 10: Vista de segmentación — Dashboard interactivo en Tableau. Fuente: elaboración propia. 44

Ilustración 11: Vista de acción comercial — Dashboard interactivo en Tableau. Fuente: elaboración propia. 45

Ilustración 12: Síntesis del proyecto — hitos y resultados principales del pipeline analítico desarrollado. Fuente: elaboración propia. 51

Índice de Tablas

Tabla 1: Configuración del modelo base XGBoost para ambos horizontes temporales. Fuente: Elaboración propia. 26

Tabla 2: Rejilla de hiperparámetros definida para la búsqueda mediante GridSearchCV. Fuente: Elaboración propia. 27

Tabla 3: Métricas globales de discriminación. Fuente: Elaboración propia. 28

Tabla 4:Métricas de evaluación Top-K empleadas en el análisis de rendimiento por percentil. Fuente: Elaboración propia. 29

Tabla 5: Herramientas tecnológicas empleadas en el proyecto. Fuente: Elaboración propia. .. 29

Tabla 6: Desglose de costes del proyecto. Fuente: Elaboración propia. 32

Tabla 7: Rendimiento del modelo base XGBoost por horizonte temporal y conjunto de evaluación. Fuente: elaboración propia. 34

Tabla 8: Hiperparámetros óptimos por horizonte temporal. Fuente: elaboración propia. 37

Tabla 9: Comparación del rendimiento entre modelo base y modelo optimizado en los cuatro conjuntos de evaluación. Fuente: elaboración propia. 37

Tabla 10: Análisis Top-K del modelo optimizado — Horizonte 6 meses, test de corte único. Fuente: elaboración propia. 40

Tabla 11: Análisis Top-K del modelo optimizado — Horizonte 3 meses, test de corte único. Fuente: elaboración propia. 40

Tabla 12: Bandas de priorización por score de propensión. Fuente: elaboración propia. 42

Tabla 13:Supuestos operativos para la estimación de impacto económico. Fuente: elaboración propia. 46

Tabla 14: Comparación de costes operativos por campaña semestral. Fuente: elaboración propia. 47

Tabla 15: Estimación de valor anual del sistema de priorización. Fuente: elaboración propia.. 48

Tabla 16: Descripción de las variables principales del dataset gold. Fuente: Elaboración propia. 57

Tabla 17: Importancia de las 20 variables principales del modelo XGBoost optimizado (horizonte de 6 meses). Fuente: Elaboración propia. 58

Tabla 18:Métricas de evaluación del modelo base y optimizado — horizonte de 6 meses. Fuente: Elaboración propia. 59

Tabla 19: Métricas de evaluación del modelo base y optimizado — horizonte de 3 meses. Fuente: Elaboración propia. 59

Tabla 20: Análisis Top-K del modelo optimizado de 6 meses sobre el test de corte único (agosto 2025, 32,344 clientes, 39 positivos observados, tasa base 0.12%). Fuente: Elaboración propia. 59

Tabla 21: Estructura de hojas del entregable comercial (entregable_marketing_scoring.xlsx). Fuente: Elaboración propia. 60

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

Nacional Seguros dispone de información histórica de contratantes naturales y sus compras de diferentes productos, lo que permite fortalecer la gestión comercial mediante el uso de datos. El proyecto se justifica por la necesidad de mejorar la efectividad de las campañas de *cross-sell*, incorporando un enfoque más sistemático y orientado a la toma de decisiones basada en datos.

1.2 Planteamiento del problema

Actualmente, las campañas de *cross-sell* se realizan de forma poco direccionada, lo que limita la capacidad de priorizar clientes con mayor potencial y reduce la eficiencia del esfuerzo comercial. En este contexto, se requiere un enfoque que permita identificar segmentos de clientes relevantes y orientar acciones comerciales de manera más efectiva, resolviendo un problema empresarial concreto.

1.3 Objetivos del proyecto

El objetivo del proyecto es segmentar a los contratantes naturales de Nacional Seguros mediante un enfoque *data-driven*, con el fin de identificar segmentos con mayor potencial para campañas direccionadas de *cross-sell* del producto Casa Protegida.

1.4 Resultados obtenidos

Se construyó un *dataset* analítico de 1,268,176 registros a partir de la base de datos corporativa, integrando variables demográficas y de comportamiento de compra de las líneas Salud, Vida y Auto. Sobre este *dataset* se entrenó un modelo supervisado de propensión basado en XGBoost, optimizado mediante búsqueda de hiperparámetros con validación cruzada cronológica. El modelo permite priorizar clientes según su probabilidad estimada de adquisición de Casa Protegida en un horizonte de seis meses.

La evaluación sobre datos no vistos durante el entrenamiento mostró que el segmento del 0.5% de clientes con mayor score concentra una proporción de compradores efectivos aproximadamente sesenta veces superior a la selección aleatoria. Como entregable operativo se generó un archivo con el ranking de 32,344 clientes organizados en seis bandas de prioridad, diseñado para su uso directo por el equipo comercial en campañas de *cross-sell* focalizadas.

1.5 Estructura de la memoria

La memoria se organiza en nueve capítulos. El Capítulo 1 presenta el contexto, el problema y los objetivos del proyecto. El Capítulo 2 revisa el estado del arte en *cross-sell*, modelos de propensión y métricas de evaluación bajo desbalance de clases. El Capítulo 3 detalla los objetivos generales, específicos y los beneficios esperados. El Capítulo 4 describe el desarrollo del proyecto: planificación, metodología, herramientas, recursos, presupuesto, viabilidad y

resultados obtenidos. El Capítulo 5 discute los resultados y las limitaciones del estudio. El Capítulo 6 recoge las conclusiones del trabajo y las conclusiones personales. El Capítulo 7 propone futuras líneas de trabajo. El Capítulo 8 lista las referencias bibliográficas y el Capítulo 9 incluye los anexos complementarios.

1.6 Alineación con los Objetivos de Desarrollo Sostenible

El presente proyecto está relacionado con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030 de las Naciones Unidas (Naciones Unidas, 2015), específicamente con el ODS 9 (Industria, innovación e infraestructura) y, complementariamente, con el ODS 8 (Trabajo decente y crecimiento económico).

La relación con el ODS 9 se concreta en la meta 9.5 que fomenta la investigación científica y el mejoramiento de la capacidad tecnológica de los sectores industriales de todos los países, en particular de los países en desarrollo. El desarrollo de un sistema de *scoring* con base en aprendizaje automático para la priorización comercial en el sector asegurador boliviano, es una aplicación directa de técnicas de ciencia de datos, en un contexto donde la adopción de estas herramientas es todavía incipiente. El proyecto muestra que se pueden implementar modelos predictivos con datos reales de nuestro propio país y con recursos computacionales accesibles, y que esto ayuda a acortar la brecha tecnológica en la gestión comercial del sector financiero nacional.

Respecto al ODS 8, el proyecto se alinea con los objetivos 8.2 y 8.10. La meta 8.2 fomenta la diversificación, la modernización tecnológica y la innovación como medios para incrementar la productividad económica. Los modelos de propensión para optimizar las campañas de *cross-sell* permiten orientar los recursos comerciales a los segmentos con mayor probabilidad de conversión, lo cual incrementa la eficiencia operativa de la aseguradora. La meta 8.10 se refiere a impulsar la capacidad de las instituciones financieras nacionales para ampliar el acceso a servicios bancarios, financieros y de seguros para todos. Al permitir identificar mejor a los clientes más afines al producto Casa Protegida, el modelo ayuda a mejorar la cobertura aseguradora del patrimonio de los hogares bolivianos.

Capítulo 2. ANTECEDENTES / ESTADO DEL ARTE

2.1 Estado del arte

La ejecución de campañas de *cross-sell* efectivas mediante enfoques *data-driven* constituye un reto analítico y operativo en el sector asegurador, debido a la heterogeneidad del comportamiento del cliente, la baja frecuencia del evento objetivo y las limitaciones de las estrategias comerciales basadas únicamente en reglas o segmentaciones estáticas (Kaishev et al., 2013; Ngai et al., 2009).

En el marco de la Customer Relationship Management (CRM), la literatura muestra que la analítica y la minería de datos se han utilizado ampliamente para mejorar procesos de adquisición, retención y expansión, mediante la explotación de datos históricos de clientes y transacciones (Ngai et al., 2009).

Desde el punto de vista metodológico, la ejecución de proyectos de minería de datos y analítica predictiva se beneficia de marcos de proceso estructurados. En este ámbito, CRISP-DM (Cross-Industry Standard Process for Data Mining) se ha consolidado como el estándar de facto para el desarrollo de proyectos de descubrimiento de conocimiento a partir de datos (Wirth & Hipp, 2000). El modelo organiza el proceso en seis fases interrelacionadas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Estas fases no se ejecutan necesariamente de forma lineal, sino que admiten iteraciones y retornos entre etapas conforme se profundiza en el problema y se obtienen nuevos hallazgos (Chapman et al., 2000).

El uso de CRISP-DM resulta importante en el presente proyecto, donde cada fase del análisis se ve condicionada por la integración de datos transaccionales provenientes de múltiples ramos de seguros, la extrema baja frecuencia del evento objetivo, y la necesidad de traducir los resultados del modelo en un entregable comercial accionable. En el presente trabajo, el proceso desarrollado sigue las fases de CRISP-DM: la definición del problema de *cross-sell* y la identificación de oportunidades comerciales corresponden a la fase de comprensión del negocio; la exploración y validación del *dataset gold*, a la fase de comprensión de los datos; las transformaciones, el tratamiento de valores atípicos y la construcción de variables, a la fase de preparación de los datos; el entrenamiento y optimización del modelo XGBoost, a la fase de modelado; la evaluación con métricas orientadas a ranking y desbalance, a la fase de evaluación; y la generación del entregable de marketing con bandas de priorización y su visualización mediante un cuadro de mando interactivo, a la fase de despliegue.

En particular, el *cross-sell* se aborda como un problema de priorización: identificar un subconjunto de clientes para maximizar resultados comerciales (conversión, utilidad esperada o eficiencia de campaña), bajo restricciones operativas.

En este sentido, la literatura sugiere métodos cuantitativos para identificar eficientemente clientes de *cross-sell* en servicios financieros. Los primeros trabajos utilizaron modelos de rasgos latentes y secuenciación de productos para estimar la propensión de compra en banca

(Kamakura et al., 1991; Li et al., 2005), mientras que contribuciones más recientes han trasladado este enfoque al sector asegurador, incluyendo la optimización integral de probabilidad de venta y rentabilidad esperada por cliente (Kaishev et al., 2013). Estos trabajos refuerzan el valor de pasar de campañas masivas a estrategias focalizadas con base en la probabilidad de compra y objetivos económicos.

Respecto al modelado predictivo, la literatura reporta que los modelos supervisados han sido ampliamente utilizados para estimar la propensión a compra en contextos de CRM y marketing directo (Ngai et al., 2009).

Un desafío transversal en este dominio es el desbalance de clases, habitual cuando el evento de interés tiene baja prevalencia. Los trabajos sobre aprendizaje con datos desbalanceados describen que la presencia de clases minoritarias y distribuciones muy sesgadas pueden degradar el rendimiento de los modelos y requieren estrategias específicas de evaluación y tratamiento (Chawla et al., 2002; Fernández et al., 2018; He & Garcia, 2009). En este trabajo, el desbalance se trata principalmente mediante la creación del conjunto de entrenamiento con múltiples cortes temporales y mediante el uso de métricas informativas para eventos raros y priorización de campañas (p. ej., PR-AUC y Lift@K) (Rosset et al., 2001; Saito & Rehmsmeier, 2015).

Además, en problemas basados en comportamiento histórico es crítico la definición de ventanas temporales: una ventana de observación para construir variables a partir del historial, y una ventana de respuesta para observar si se da el evento objetivo. La longitud de la ventana de observación influye mucho en la capacidad predictiva del modelo (Ballings & Van den Poel, 2012) y la correcta separación respecto a la ventana de respuesta evita fuga de información (*data leakage*) y permite evaluar el modelo en condiciones realistas. Además, la integración y calidad de datos condicionan la disponibilidad de una perspectiva consistente del cliente y, por lo tanto, la aplicación de la analítica a escala (Kitchens et al., 2018).

Friedman (2001) formalizó el concepto de *gradient boosting*, proponiendo un marco general de aproximación funcional *greedy* mediante el cual se construye un modelo predictivo de forma aditiva mediante la incorporación secuencial de modelos débiles, árboles de decisión, ajustados al gradiente negativo de la función de pérdida. Esta estrategia admite funciones de pérdida cualesquiera sean diferenciables, lo que hace que el método sea muy flexible. XGBoost (Extreme *Gradient Boosting*), introducido por Chen y Guestrin (2016), extiende este marco con mejoras de rendimiento computacional, incluso procesamiento paralelo a nivel de variables, manejo nativo de valores faltantes y regularización (L1 y L2), lo que lo ha convertido en uno de los algoritmos más populares en competiciones y aplicaciones industriales de aprendizaje automático con datos tabulares.

La literatura ha demostrado que, en lo que se refiere a la evaluación de modelos en contextos de desbalance extremo, las métricas convencionales —tales como la exactitud (accuracy) o el área bajo la curva ROC (ROC-AUC)— pueden ser engañosas cuando la clase positiva representa una fracción muy pequeña de la población. Davis y Goadrich (2006) formalizaron la relación

entre las curvas ROC y Precision-Recall, y mostraron que un clasificador puede tener un ROC-AUC alto, pero un bajo rendimiento en precisión y recall de la clase minoritaria. Esta línea de trabajo fue reforzada por Saito y Rehmsmeier (2015), quienes confirmaron que la curva Precision-Recall es más informativa que la curva ROC para evaluar los clasificadores binarios cuando los datasets presentan distribuciones de clase altamente desbalanceadas. Ambos trabajos fundamentan la decisión de utilizar PR-AUC como la métrica principal en el presente proyecto (Véase también Fawcett, 2006).

Complementariamente, en aplicaciones de marketing directo y *propensity scoring*, la evaluación del modelo se centra con frecuencia en métricas orientadas a ranking como el *Lift* y la precisión en los percentiles superiores (Top-K). El *Lift* mide cuántas veces mejora la tasa de respuesta del modelo respecto a una selección aleatoria de la misma proporción de clientes, lo que permite cuantificar directamente el valor operativo del modelo para la priorización comercial. Rosset et al. (2001) propusieron evaluar modelos predictivos para campañas de marketing mediante métricas de ranking en lugar de métricas de clasificación pura, argumentando que lo relevante no es si el modelo clasifica correctamente a todos los individuos, sino si ordena correctamente a los más propensos en la parte superior del ranking. Este principio resulta especialmente pertinente cuando la acción comercial solo puede ejecutarse sobre un subconjunto reducido de la cartera, como ocurre en el presente proyecto con las bandas de priorización (Top 0.5%, 1%, 2% y 5%) (véase también Ling & Li, 1998; Provost & Fawcett, 2001).

En síntesis, el estado del arte sugiere que el *cross-sell* efectivo se beneficia de modelos supervisados de propensión con buen rendimiento predictivo, estrategias explícitas para tratar desbalance de clases y un diseño temporal adecuado del *dataset*, todo ello condicionado por desafíos de datos en entornos empresariales reales.

2.2 Contexto y justificación

La iniciativa se desarrolla en Nacional Seguros, empresa del sector asegurador boliviano, con una sólida cartera de clientes y un portafolio de productos diversificado. El producto Casa Protegida muestra una penetración menor en comparación con otros productos del portafolio, debido a su relativamente reciente lanzamiento, lo que reduce la disponibilidad de datos históricos y limita su adopción inicial. Sin embargo, la proporción de clientes activos que aún no cuentan con este producto indica que existe un amplio margen para campañas de venta cruzada dirigidas.

Se considera que el producto posee un valor estratégico y se le asigna prioridad en su comercialización mediante campañas de venta cruzada focalizadas. La meta es mejorar la eficiencia comercial, sustituyendo las acciones masivas de baja eficacia por esfuerzos dirigidos a los clientes con mayor probabilidad de contratación.

La contribución principal del proyecto consiste en ofrecer un método práctico y replicable para respaldar decisiones empresariales basadas en datos, que abarca desde la construcción del

conjunto de datos y el entrenamiento de un modelo de propensión, hasta la evaluación mediante métricas alineadas con la priorización de campañas de venta cruzada.

2.3 Planteamiento del problema

El estado del arte muestra que las campañas de *cross-sell* pueden beneficiarse de enfoques predictivos orientados a priorización de clientes. Sin embargo, en entornos reales del sector asegurador persisten limitaciones prácticas que dificultan su adopción sistemática.

En particular, cuando el producto objetivo tiene baja penetración, el evento de interés es infrecuente y los datos presentan un desbalance de clases marcado. En estas condiciones, las estrategias comerciales basadas únicamente en reglas o segmentaciones estáticas tienden a ser ineficientes, y la construcción de modelos predictivos se ve limitada por la escasez de casos positivos, afectando el entrenamiento y la evaluación.

En consecuencia, el problema que aborda este trabajo se centra en la falta de una solución *data-driven* que permita, a partir de los datos históricos disponibles, segmentar y caracterizar perfiles de clientes y, complementariamente, estimar la propensión de contratación del producto Casa Protegida para generar una priorización útil de clientes para campañas.

Capítulo 3. OBJETIVOS

3.1 Objetivos generales

Aplicar un enfoque analítico basado en aprendizaje supervisado sobre el histórico de ventas, comportamiento de compra y variables demográficas de los contratantes naturales de Nacional Seguros, con el fin de estimar la propensión al *cross-sell* del producto Casa Protegida mediante un modelo de *gradient boosting* (XGBoost), generando un ranking de priorización y perfiles de clientes por bandas de *scoring* orientados a campañas comerciales focalizadas, y visualizando los resultados mediante dashboards en Tableau para apoyar la toma de decisiones.

3.2 Objetivos específicos

- Consolidar el histórico de ventas de contratantes naturales de Nacional Seguros, integrando variables demográficas y de comportamiento de compra (edad, género, antigüedad, cantidad de pólizas, anulaciones y primas pagadas) para los productos Automotores, Salud, Vida largo plazo y Casa Protegida.
- Preparar y transformar el conjunto de datos para el análisis, incluyendo limpieza, tratamiento de valores faltantes y atípicos, y estandarización/normalización de variables, garantizando consistencia y comparabilidad.
- Definir y seleccionar las variables explicativas representativas del comportamiento de compra y del perfil demográfico, reservando la variable “contratación de Casa Protegida” como variable objetivo (*target*) del modelo supervisado.
- Desarrollar un modelo supervisado de propensión al *cross-sell* basado en XGBoost, con diseño temporal de entrenamiento y evaluación que asegure la ausencia de fuga de información, para estimar la probabilidad de adquisición del producto Casa Protegida en un horizonte de seis meses.
- Evaluar el modelo mediante métricas orientadas a priorización comercial bajo condiciones de desbalance extremo de clases, incluyendo PR-AUC, Lift@K y Precision@K, e identificar las variables con mayor influencia en la predicción mediante la importancia de variables del modelo.
- Generar un entregable comercial basado en el *scoring* de propensión, con un ranking de priorización por bandas y perfiles de clientes, orientado a facilitar la ejecución de campañas de *cross-sell* focalizadas sobre subconjuntos top-k de clientes.
- Visualizar los resultados analíticos y de negocio mediante dashboards en Tableau, orientados a apoyar la toma de decisiones comerciales y el seguimiento de campañas de *cross-sell*.

3.3 Beneficios del proyecto

El proyecto aporta a Nacional Seguros una metodología *data-driven* para apoyar campañas de *cross-sell*, basada en el análisis de datos históricos, variables demográficas y comportamiento de compra. Esta metodología permite priorizar contratantes con mayor propensión a la contratación del producto Casa Protegida, facilitando la focalización de esfuerzos comerciales y la planificación de campañas.

El *scoring* de propensión con XGBoost, complementado con métricas de priorización (Top-K y *Lift*) y visualización en Tableau, habilita un proceso replicable y medible para la toma de decisiones. Su aplicación contribuye a mejorar la eficiencia operativa de las campañas, reduciendo acciones masivas y orientando recursos hacia perfiles con mayor potencial de contratación.

Adicionalmente, tanto el *pipeline* de extracción y transformación de datos (SQL) como el *pipeline* analítico de modelado y evaluación han sido diseñados de forma modular y documentada, de modo que pueden reutilizarse para otros productos del portafolio ajustando las variables explicativas y la variable objetivo. Esta característica extiende el valor del proyecto más allá del caso específico de Casa Protegida, proporcionando una herramienta analítica aplicable a iniciativas futuras de *cross-sell* o retención.

Para cuantificar el valor monetario aportado por el modelo en futuras campañas de *cross-sell*, se propone una metodología basada en el ingreso incremental esperado. Este se calcula como la diferencia entre la tasa de conversión observada al contactar al Top-K priorizado por el *scoring* y la tasa de conversión bajo selección aleatoria, multiplicada por el número de contratantes contactados y por la prima promedio del producto Casa Protegida. La expresión formal es la siguiente:

$$Vincremental = (TTop - K - Tbase) \times Ncontactos \times Ppromedio$$

Ecuación 1 Estimación del ingreso incremental por priorización Top-K

donde $T_{Top\text{-}K}$corresponde a la tasa de conversión del grupo priorizado por el modelo, T_{base}a la tasa de conversión bajo selección aleatoria, $N_{contactos}$al número de contratantes contactados en la campaña y $P_{promedio}$a la prima promedio del producto Casa Protegida. El *Lift* obtenido en el conjunto de validación permite anticipar el orden de magnitud del beneficio, mientras que su cuantificación definitiva requerirá datos de conversión real provenientes de una campaña piloto posterior al despliegue del modelo.

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

El proyecto se ha desarrollado entre octubre de 2025 y abril de 2026, con una dedicación parcial de aproximadamente cinco a diez horas semanales, compaginando el trabajo del TFM con la actividad profesional del autor. A continuación, se describen las actividades principales y su distribución temporal.

En octubre de 2025 se realizó la definición del problema de negocio, la revisión de fuentes de datos corporativas disponibles y el diseño del esquema temporal de cortes, incluyendo la definición de las ventanas de observación y de respuesta. Durante noviembre de 2025 se desarrollaron las consultas SQL específicas por línea de negocio (Automotores, Salud y Vida), incorporando las particularidades de cada producto, y se construyó el *dataset* consolidado a partir de la información extraída de SQL Server.

Entre diciembre de 2025 y febrero de 2026 se ejecutó el desarrollo analítico en Python 3.13.2 mediante JupyterLab dentro de un entorno Anaconda: carga y validación inicial del *dataset* (diciembre), análisis exploratorio y preparación de datos para modelado (diciembre–enero), entrenamiento del modelo base XGBoost y optimización de hiperparámetros mediante GridSearchCV (enero–febrero), y generación del entregable de marketing con *scoring* y bandas de priorización (febrero).

A partir de enero de 2026 se inició la redacción de la memoria en paralelo con el desarrollo técnico. Durante marzo y abril de 2026 se completó la redacción, se desarrollaron los dashboards en Tableau y se realizaron las revisiones con el director del proyecto, Dr. Alberto Partida. De forma transversal, a lo largo de todo el periodo se realizaron reuniones periódicas de seguimiento con el tutor.

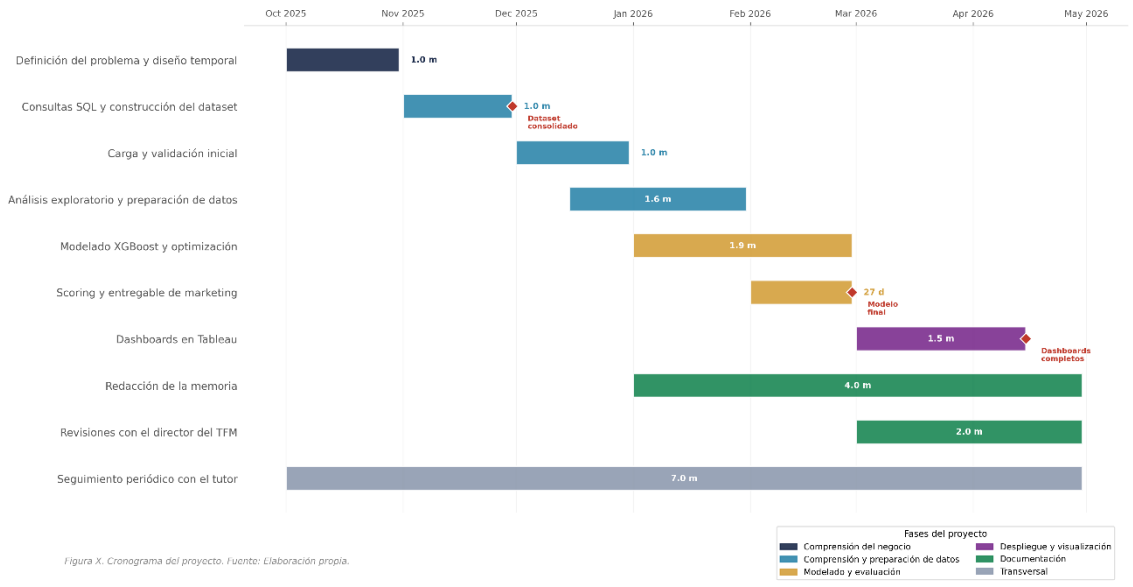

Ilustración 1: Cronograma del proyecto. Fuente: Elaboración propia

4.2 Descripción de la solución, metodologías y herramientas empleadas

4.2.1 Visión general de la solución

Se ha desarrollado una solución analítica orientada a la venta cruzada del producto Casa Protegida, basada en la probabilidad de contratación a nivel de contratante. La solución produce un score por contratante y un ranking de priorización para ejecutar acciones comerciales sobre un subconjunto top-k.

El desarrollo se apoya en la construcción de un *dataset* analítico reproducible, con un diseño temporal de cortes que permite generar múltiples observaciones por contratante en distintos momentos, incrementando el número de ejemplos positivos disponibles para el entrenamiento. Sobre esta base se desarrolla un modelo supervisado de propensión mediante XGBoost, optimizado con búsqueda de hiperparámetros y evaluado con métricas orientadas a priorización comercial en escenarios de desbalance extremo de clases.

4.2.2 Metodología de desarrollo

El proyecto se ha desarrollado de forma iterativa. La iteración se guía por criterios de coherencia temporal (FechaCorte), por la disponibilidad de información en las fuentes y por el rendimiento del modelo en métricas orientadas a priorización comercial.

El proceso ejecutado se estructura en las siguientes etapas:

1) Definición del esquema temporal de análisis, estableciendo una FechaCorte por iteración y las ventanas de cálculo asociadas (histórico para variables y ventana futura para identificación del evento objetivo).
2) Construcción de consultas específicas por línea de negocio (Auto, Salud y Vida) para derivar variables de comportamiento, incorporando particularidades por producto (por ejemplo, cantidad de vehículos en Auto y cantidad de asegurados en Salud).
3) Cálculo de variables transversales por producto, incluyendo recencia de compra, antigüedad desde la primera compra, y normalización de importes a una moneda común en dólares estadounidenses.
4) Aplicación de criterios de vigencia a FechaCorte en cada consulta para definir, por corte, el universo elegible y asegurar consistencia temporal.
5) Consolidación del universo por corte y enriquecimiento mediante atributos del contratante (por ejemplo, fecha de nacimiento, género y otros atributos disponibles en fuentes corporativas).
6) Almacenamiento del resultado por corte en un *dataset* consolidado final, utilizado como base para análisis y modelado.
7) Entrenamiento de un modelo supervisado de propensión basado en XGBoost, utilizando las variables explicativas derivadas del *dataset gold*, con pesos temporales para priorizar la representatividad de los cortes más recientes.
8) Optimización de hiperparámetros mediante búsqueda en grilla (GridSearchCV) con validación cruzada cronológica interna, y evaluación final con métricas orientadas a

ranking (top-k) y a escenarios desbalanceados (PR-AUC), sobre un conjunto de prueba temporalmente separado del entrenamiento.

Dado que en un solo corte temporal se observa una cantidad muy baja de casos positivos (contrataciones de Casa Protegida), se definió la estrategia de combinar varios cortes temporales consistentes. Con ello se incrementa el número de ejemplos positivos para el entrenamiento del modelo, manteniendo la coherencia temporal del análisis.

4.2.3 Diseño del análisis de datos

El análisis se estructura sobre un diseño temporal de cortes mensuales. Cada corte, identificado por una FechaCorte, define un momento de observación a partir del cual se construyen las variables explicativas y se identifica el evento objetivo. Para cada FechaCorte se establecen dos ventanas: una ventana de observación de doce meses hacia atrás, utilizada para calcular las variables de comportamiento de contratación por línea de negocio, y una ventana de respuesta de seis meses hacia adelante, utilizada para determinar si el contratante adquirió el producto Casa Protegida en dicho horizonte.

Esta estructura genera un registro por contratante y por corte, con lo cual un mismo contratante puede aparecer en múltiples cortes con distintos valores en sus variables y en su *target*. El *dataset* resultante comprende 1,268,176 registros, distribuidos en 41 fechas de corte entre junio de 2020 y noviembre de 2025.

Se definieron dos variables objetivo, correspondientes a horizontes temporales distintos, con el fin de evaluar durante la fase de modelado cuál ofrece un mejor equilibrio entre cantidad de ejemplos positivos y capacidad predictiva. La variable cp_compro_3m toma valor 1 si el contratante contrató al menos una póliza de Casa Protegida dentro de los tres meses posteriores a la FechaCorte, y 0 en caso contrario. La variable cp_compro_6m aplica la misma lógica para un horizonte de seis meses. Ambas presentan un desbalance extremo: cp_compro_3m registra 425 casos positivos (tasa base de 0,034%) y cp_compro_6m registra 809 casos positivos (tasa base de 0,064%), sobre un total de 1,268,176 registros. La selección definitiva del horizonte se realiza en función de los resultados de entrenamiento y evaluación del modelo.

El *dataset* consolidado contiene 46 columnas organizadas en las siguientes familias: variables demográficas del contratante (edad, género, estado civil), *flags* de portafolio (indicadores de tenencia de pólizas activas en Salud, Vida y Automotores en los últimos doce meses), variables de comportamiento por línea de negocio (cantidad de pólizas, montos en dólares, renovaciones, anulaciones, recencia y antigüedad), y las variables objetivo para los horizontes de tres y seis meses. La separación entre variables explicativas y variables objetivo se aplica de forma estricta para evitar fuga de información.

4.2.3.1 Ingeniería de variables por línea de negocio

Para armar el *dataset* analítico se han desarrollado consultas específicas por línea de negocio (Automotores, Salud y Vida), incorporando variables de comportamiento calculadas sobre una

ventana de doce meses previos a cada FechaCorte. Las variables dependientes se agrupan en las siguientes agrupaciones:

En Automotores se consideran variables de volumen de aseguramiento, por ejemplo, cantidad de vehículos asegurados, variables de comportamiento de cartera, tales como renovaciones y anulaciones, y variables monetarias tales como prima.

En Salud se consideran variables de volumen de aseguramiento que dependen del número de asegurados y variables de comportamiento de cartera, tales como renovaciones y anulaciones, variables monetarias tales como prima y variables de comportamiento de pago.

En Vida se consideran variables centradas en pagos como la cantidad de cuotas pagadas en los últimos doce meses y la fecha del último pago, así como variables monetarias y de comportamiento relacionadas con el producto.

Por otra parte, para cada línea de negocio se calculan métricas temporales como recencia, definida como el tiempo transcurrido desde la última contratación o pago, y antigüedad, definida como el tiempo desde la primera contratación. También, para asegurar la comparabilidad, los montos monetarios se expresan en una única moneda: dólares estadounidenses.

4.2.3.2 Coherencia temporal y definición de universo por corte

En cada consulta por producto se aplican filtros de fechas y criterios de vigencia para asegurar que, para cada FechaCorte, únicamente se consideren registros vigentes a dicha fecha. Esta regla permite definir el universo elegible de contratantes por corte y evita incorporar información posterior al momento de predicción.

4.2.3.3 Enriquecimiento de atributos del contratante y consolidación del dataset

A partir de los identificadores de contratante obtenidos en las consultas por producto, se incorporan atributos adicionales del contratante, tales como fecha de nacimiento, género y otros atributos disponibles en las fuentes corporativas. Finalmente, se consolida la información por corte en una tabla final que constituye el *dataset* consolidado utilizado para las etapas posteriores de análisis y modelado. La descripción de resultados exploratorios, prevalencia del *target* y comportamiento de variables se presenta en la sección de resultados.

La Ilustración 2 presenta de forma esquemática el *pipeline* de extracción y consolidación implementado en SQL Server, desde las tablas fuente de la base de datos corporativa hasta la obtención del *Dataset* Gold exportado en formato Parquet.

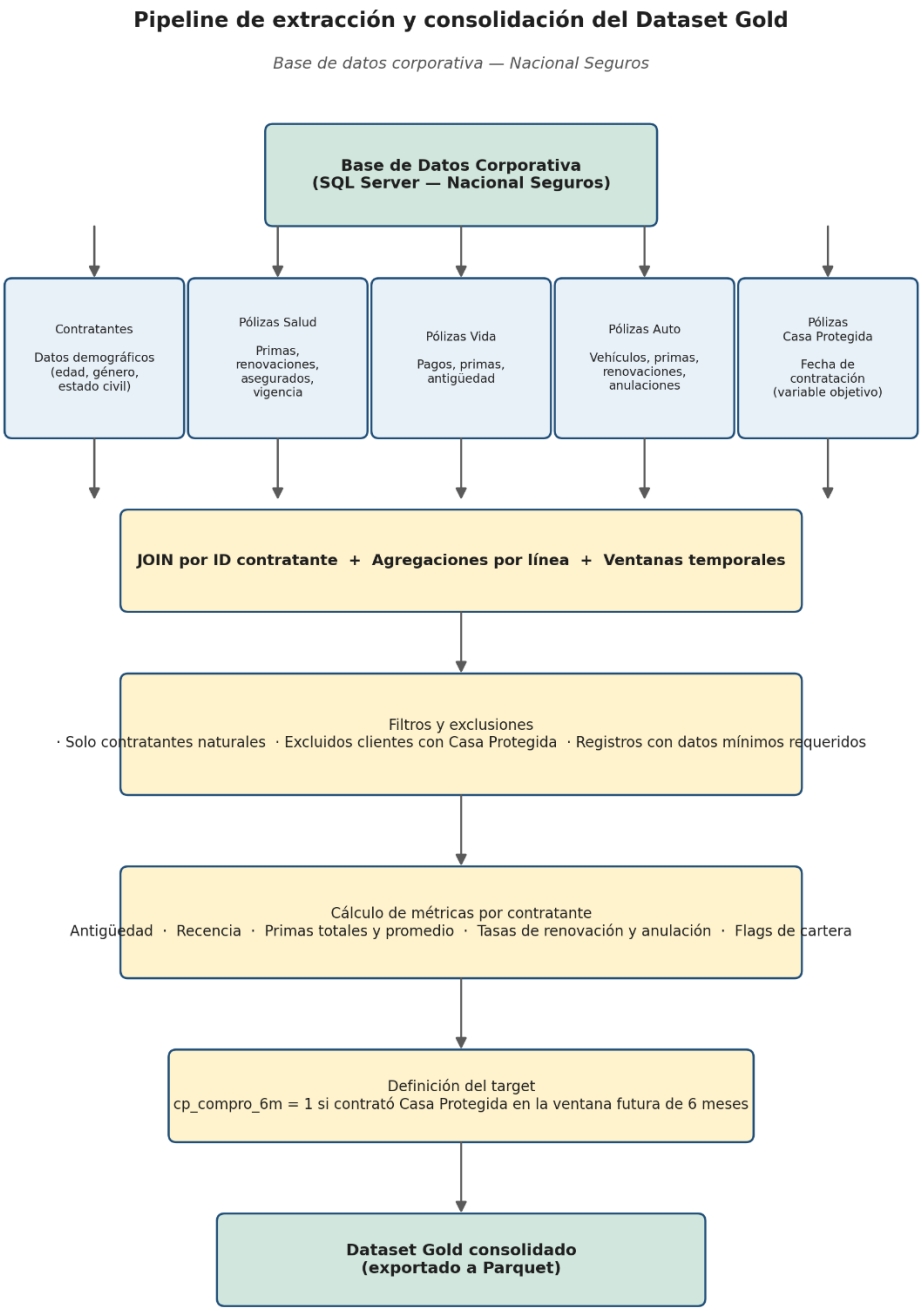

Ilustración 2:Pipeline de extracción y consolidación del Dataset Gold. Fuente: elaboración propia.

4.2.4 Modelos de prospección

La Ilustración 3 sintetiza las etapas del *pipeline* analítico implementado en JupyterLab, desde la carga del *Dataset* Gold hasta la generación de los entregables finales del proyecto.

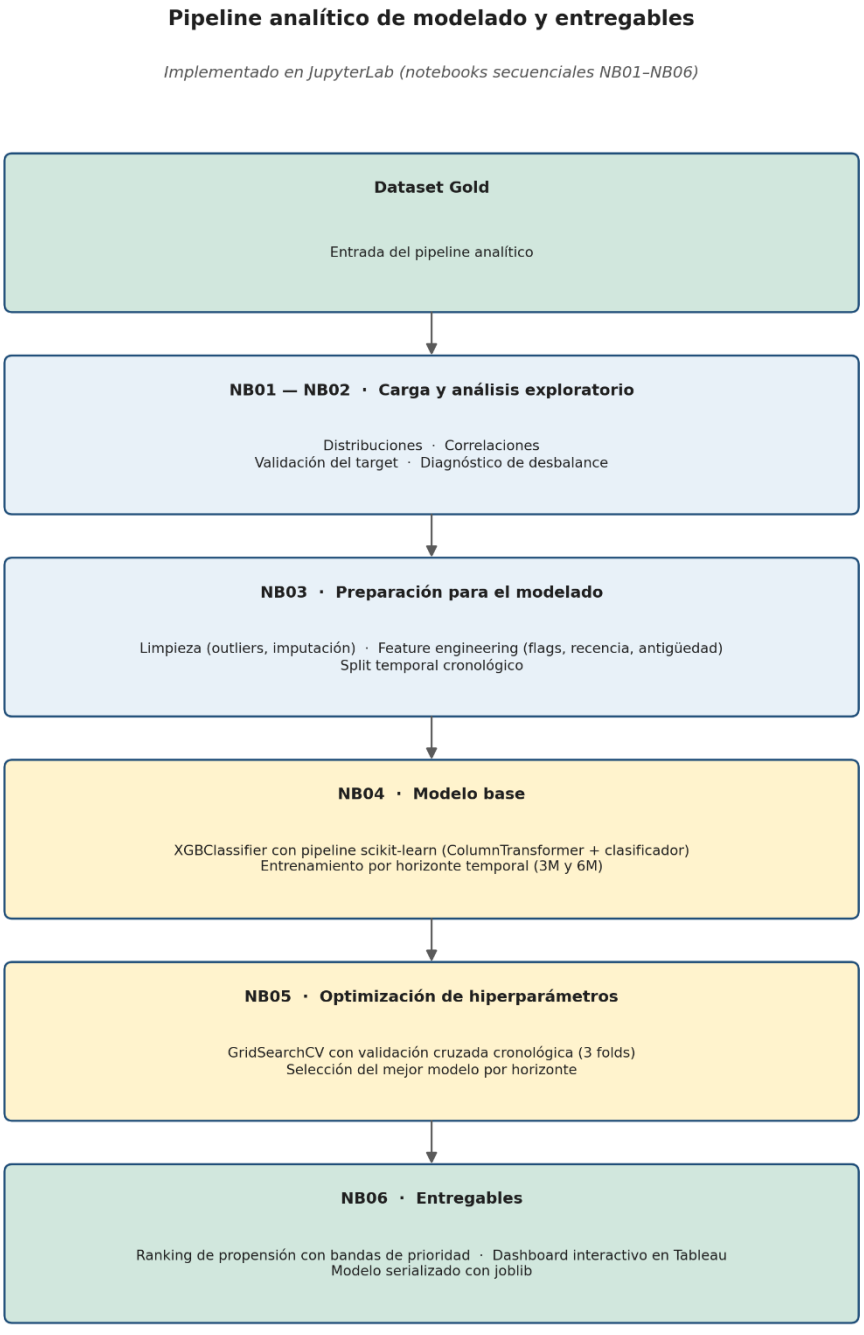

Ilustración 3:Pipeline analítico de modelado y entregables. Fuente: elaboración propia.

El núcleo del sistema de *scoring* está basado en un modelo supervisado de clasificación binaria que tiene como objetivo estimar la probabilidad de que un contratante activo contrate el producto Casa Protegida en un horizonte futuro definido. Dado el carácter tabular del conjunto de datos y el desequilibrio extremo de clases, se escogió XGBoost (Extreme *Gradient Boosting*) como algoritmo principal, por su robustez frente a datos heterogéneos, su manejo nativo de valores faltantes y su eficacia contrastada en problemas de clasificación con distribuciones muy asimétricas de clase.

El modelo se implementó mediante un *pipeline* unificado de scikit-learn que encapsula las etapas de preprocesamiento y clasificación en un único objeto. El preprocesamiento se estableció mediante un ColumnTransformer con dos rutas paralelas: imputación por mediana (SimpleImputer(strategy="median")) para las variables numéricas, e imputación por moda (SimpleImputer(strategy="most_frequent")) seguida de codificación one-hot (OneHotEncoder(handle_unknown="ignore")) para las variables categóricas. Ambas ramas entran al clasificador XGBClassifier, conformando un *pipeline end-to-end* que asegura la reproducibilidad y evita fugas de información entre los *splits* de entrenamiento y evaluación.

Se construyeron dos modelos con la misma arquitectura, uno para cada horizonte temporal definido (cp_compro_3m y cp_compro_6m), diferencian en la variable objetivo y en el valor de scale_pos_weight. Este parámetro se configuró con el cociente entre observaciones negativas y positivas en el conjunto de entrenamiento para abordar el desbalance de clases: 3 518 para el *target* a tres meses y 2 528 para el *target* a seis meses, lo que indica al algoritmo la ponderación que debe asignar a cada instancia positiva durante el entrenamiento. Los dos modelos base se ajustaron con 500 árboles (n_estimators=500), una tasa de aprendizaje de 0.05 (learning_rate=0.05) y una profundidad máxima de 5 niveles (max_depth=5).

Parámetro	Valor 3m	Valor 6m	Descripción
n_estimators	500	500	Número de árboles
learning_rate	0.05	0.05	Tasa de aprendizaje
max_depth	5	5	Profundidad máxima por árbol
scale_pos_weight	3 518	2 528	Ponderación de clase positiva
eval_metric	aucpr	aucpr	Métrica de monitoreo interno

Tabla 1: Configuración del modelo base XGBoost para ambos horizontes temporales. Fuente: Elaboración propia.

La partición temporal entre conjunto de entrenamiento y conjunto de prueba se hizo siguiendo un orden estricto cronológico: los cortes más antiguos fueron usados para entrenamiento y los más recientes para test, incorporando además un periodo de separación (gap) entre ambos conjuntos para evitar contaminación temporal. Además, se aplicó una ponderación temporal sobre las observaciones de entrenamiento en tres bandas, asignando un peso de 1.0 para los cortes hasta diciembre de 2023, un peso de 1.5 para los cortes entre enero y agosto de 2024, y

un peso de 2.0 para los cortes entre septiembre de 2024 y mayo de 2025, en función de la premisa de que los patrones de comportamiento más cercanos al presente son más representativos de la propensión futura.

Se estableció la línea base y se procedió a la optimización de hiperparámetros mediante GridSearchCV con validación cruzada cronológica, que se ejecutó de forma independiente para cada horizonte temporal. Se estableció una malla de siete hiperparámetros, cada uno con dos posibles valores, lo que dio lugar a un total de 128 combinaciones (2^7). Los hiperparámetros explorados fueron: número de árboles (n_estimators: 300, 500), tasa de aprendizaje (learning_rate: 0.03 y 0.05), profundidad máxima (max_depth: 3 y 5), peso mínimo por hoja (min_child_weight: 1 y 5), proporción de muestreo por árbol (subsample: 0.8 y 1.0), proporción de columnas por árbol (colsample_bytree: 0.8 y 1.0) y regularización L2 (reg_lambda: 1.0 y 5.0).

Hiperparámetro	Valores candidatos	Descripción
n_estimators	300 / 500	Número de árboles
learning_rate	0.03 / 0.05	Tasa de aprendizaje
max_depth	3 / 5	Profundidad máxima por árbol
min_child_weight	1 / 5	Peso mínimo por hoja
subsample	0.8 / 1.0	Proporción de muestreo por árbol
colsample_bytree	0.8 / 1.0	Proporción de columnas por árbol
reg_lambda	1.0 / 5.0	Regularización L2

Tabla 2: Rejilla de hiperparámetros definida para la búsqueda mediante GridSearchCV. Fuente: Elaboración propia.

Para la estrategia de validación cruzada se usó PredefinedSplit de scikit-learn, que permite establecer de forma explícita qué observaciones pertenecen a cada *fold* de validación. Se establecieron tres *folds* cronológicos, en los que cada partición de entrenamiento va antes, en el tiempo, que su partición de validación correspondiente. Esta estrategia respeta la secuencia natural de los datos y evita que aparezcan datos del futuro en el entrenamiento, lo que es fundamental en problemas con dependencia temporal. Se utilizó average_precision (equivalente al área bajo la curva Precision-Recall, PR-AUC) como métrica de selección, debido a su sensibilidad al desempeño del modelo sobre la clase minoritaria en contextos de desbalance extremo.

La mejor configuración que cada búsqueda de parámetros arroja se reporta en el capítulo de resultados, junto con el análisis comparativo de desempeño entre ambos horizontes temporales. Se utilizó joblib para serializar todo el *pipeline*, tanto el preprocesamiento como el modelo optimizado, y así poder utilizarlo de forma portátil y reutilizable durante la fase de *scoring* sobre nuevos datos.

4.2.5 Diseño de validación y métricas

La evaluación del modelo de propensión se organizó en dos niveles complementarios: métricas globales de discriminación y métricas de rendimiento en los percentiles superiores del ranking (Top-K). Esta doble perspectiva responde a la naturaleza del problema de negocio, donde no se trata de obtener la correcta clasificación de toda la cartera, sino de identificar con precisión el subconjunto de clientes con mayor probabilidad de contratación con el fin de priorizar la acción comercial.

El área bajo la curva Precision-Recall (PR-AUC), también conocida como average_precision en scikit-learn, se usó como la métrica principal de selección. En situaciones como la actual, de desbalance extremo (tasa base del 0,064%), la PR-AUC es más informativa que la exactitud o el área bajo la curva ROC, ya que esta última puede dar valores aparentemente altos incluso si el modelo es incapaz de detectar la clase minoritaria. La PR-AUC, por el contrario, penaliza los falsos positivos en proporción a la prevalencia real, ofreciendo una imagen más fiel del rendimiento operativo del modelo. Sin embargo, se incluye el ROC-AUC como métrica secundaria de referencia para facilitar la comparabilidad.

Métrica	Rol en la evaluación	Justificación
PR-AUC	Métrica principal de selección	Sensible al rendimiento sobre la clase minoritaria en desbalance extremo
ROC-AUC	Métrica secundaria de referencia	Facilita comparabilidad con otros estudios

Tabla 3: Métricas globales de discriminación. Fuente: Elaboración propia.

El segundo nivel de evaluación consistió en un análisis Top-K con el fin de medir el rendimiento del modelo en los segmentos más prioritarios del ranking de probabilidades. Se consideraron cinco percentiles: Top 1%, Top 3%, Top 5%, Top 10% y Top 20% de la cartera ordenada por score descendente. Para cada percentil se calcularon tres indicadores: Precision@K (proporción de positivos reales dentro del Top-K), Recall@K (proporción del total de positivos recuperados dentro del Top-K) y Lift@K (relación entre la Precision@K y la tasa base, que mide cuántas veces mejora el modelo con respecto a una selección aleatoria). El *Lift* es la métrica con mayor relevancia operativa, dado que mide directamente la ganancia de eficiencia que el equipo comercial obtiene al utilizar el modelo comparado con un enfoque no dirigido.

Métrica	Definición	Interpretación
Precision@K	Proporción de positivos reales dentro del Top-K del ranking	Qué tan preciso es el modelo al seleccionar los K clientes más prioritarios
Recall@K	Proporción del total de positivos capturados dentro del Top-K	Qué cobertura del total de compradores potenciales se alcanza con el Top-K

Lift@K	Precision@K / tasa base	Cuántas veces mejora el modelo respecto a una selección aleatoria

Tabla 4:Métricas de evaluación Top-K empleadas en el análisis de rendimiento por percentil. Fuente: Elaboración propia.

La evaluación final del modelo optimizado se realizó en un conjunto de prueba totalmente independiente, conformado por los cortes temporales más recientes, que no fueron utilizados ni en el entrenamiento ni en la validación cruzada. Se dan las métricas tanto sobre el conjunto de prueba completo como sobre el último corte individual para comprobar la estabilidad del desempeño en el periodo más cercano al presente. En la sección de resultados, se detallan los resultados de ambos niveles de evaluación.

4.2.6 Herramientas tecnológicas y recursos de cálculo

El desarrollo del proyecto se apoyó en un conjunto de herramientas de software libre y propietario que cubren las distintas fases del proceso analítico. La Tabla 5 resume las herramientas empleadas, organizadas por fase del proyecto.

Categoría	Herramienta	Función
Lenguaje y entorno	Python (3.13.2)	Lenguaje de programación principal
	Anaconda	Gestión de paquetes y dependencias
	JupyterLab	Desarrollo interactivo en *notebooks*
Base de datos	SQL Server	Motor de base de datos corporativo
Procesamiento de datos	pandas	Transformación y análisis tabular
	numpy	Operaciones numéricas
	Apache Parquet	Formato columnar de almacenamiento intermedio
Aprendizaje automático	scikit-learn	Preprocesamiento, pipelines, validación y búsqueda de hiperparámetros
	XGBoost	Algoritmo de clasificación (*gradient boosting*)
Serialización	joblib	Persistencia de pipelines entrenados
Visualización y reportería	matplotlib	Gráficos de evaluación del modelo
	openpyxl	Motor de exportación Excel (vía pandas ExcelWriter)
Business Intelligence	Tableau Desktop	Dashboards interactivos de negocio
Utilidades	pathlib	Portabilidad de rutas entre sistemas operativos

Tabla 5: Herramientas tecnológicas empleadas en el proyecto. Fuente: Elaboración propia.

El lenguaje de programación utilizado fue Python en su versión 3.13.2, ejecutado sobre la distribución Anaconda, la cual provee un entorno gestionado de paquetes y dependencias. Para

el entorno de desarrollo interactivo se empleó JupyterLab, que permite combinar código, visualizaciones y documentación en *notebooks* que pueden reproducirse.

Para la manipulación y transformación de datos se utilizaron las librerías pandas y numpy. El *dataset* consolidado, que incorpora información de diversas líneas de negocio, fue construido a través de consultas SQL ejecutadas sobre el motor de base de datos corporativo SQL Server de Nacional Seguros. Los resultados intermedios entre etapas del *pipeline* fueron almacenados en formato Apache Parquet, un formato columnar binario que brinda compresión eficiente y tiempos de lectura significativamente menores que los formatos tabulares tradicionales como CSV.

En la fase de modelado se empleó scikit-learn como el principal framework de aprendizaje automático, haciendo uso de sus módulos de preprocesamiento (SimpleImputer, OneHotEncoder, ColumnTransformer), composición de pipelines (*Pipeline*), partición temporal (PredefinedSplit) y búsqueda de hiperparámetros (GridSearchCV). Se escogió como algoritmo de clasificación el XGBoost (XGBClassifier), una implementación de *gradient boosting* optimizada para rendimiento y escalabilidad. Se realizó la serialización de los pipelines entrenados con joblib para poder persistir y cargar los modelos sin necesidad de volverlos a entrenar.

Se utilizó matplotlib para la generación de los gráficos de evaluación: curvas ROC, curvas de Precision-Recall y matrices de confusión. El entregable comercial final fue un archivo Excel con varias hojas que incluye el ranking de clientes y sus bandas de priorización. Se generó en openpyxl. Para gestionar las rutas y la estructura de directorios del proyecto se ha usado pathlib, lo cual garantiza que el código sea portable entre diferentes sistemas operativos.

El código fuente completo del proyecto se encuentra disponible en un repositorio público de GitHub [starifa85/TFM_Cross_Sell_Casa_Protegida: Trabajo Final de Máster: modelo predictivo de cross-sell para el producto Casa Protegida, aplicando técnicas de machine learning sobre datos de clientes de seguros](), organizado en *notebooks* secuenciales que cubren desde la preparación de datos hasta la generación del entregable comercial, lo cual permite la reproducibilidad íntegra del proceso analítico.

Por otro lado, se empleó Tableau Desktop como herramienta de Business Intelligence para generar cuadros de mando interactivos que permiten al equipo comercial explorar los resultados del modelo, filtrar por segmentos y observar la distribución de propensiones a lo largo de la cartera. Todo el desarrollo se realizó en la máquina personal del autor, sin necesidad de infraestructura de cómputo adicional más allá del acceso a la base de datos corporativa.

4.3 Recursos requeridos

Para la ejecución del presente proyecto se ha requerido un conjunto de recursos de naturaleza diversa. La descripción detallada de las herramientas de software puede consultarse en el apartado 4.2.6.

En cuanto a recursos de hardware, se ha utilizado un ordenador portátil personal equipado con un procesador Intel Core i7-1360P de 13.ª generación (2.20 GHz), 16 GB de memoria RAM (6000

MT/s) y un disco de estado sólido (SSD) de 477 GB, operando bajo Windows 11 Home. No se ha requerido infraestructura de cómputo adicional ni procesamiento en GPU, dado que el volumen de datos (1.27 millones de registros, 46 variables) y la complejidad computacional del modelo —incluyendo la búsqueda de 128 combinaciones de hiperparámetros con validación cruzada de 3 folds— resultaron manejables con esta configuración.

Respecto al software, el proyecto se ha desarrollado íntegramente con herramientas de código abierto (Python, scikit-learn, XGBoost, entre otras), con la única excepción de Tableau Desktop, que opera bajo licencia comercial y fue empleada para la elaboración de dashboards de visualización de resultados.

En cuanto a infraestructura de datos, se ha contado con acceso autorizado a la base de datos corporativa de Nacional Seguros (motor SQL Server), desde la cual se extrajeron las tablas necesarias para la construcción del *dataset* consolidado.

Finalmente, en lo que respecta a recursos humanos, el proyecto ha contado con la dedicación del autor, con una carga estimada de entre 220 y 260 horas distribuidas a lo largo de siete meses, y con la dirección académica del Dr. Alberto Partida como tutor del Trabajo Fin de Máster.

4.4 Presupuesto

El presupuesto del presente proyecto refleja la evaluación económica de todos los recursos empleados durante su ejecución. Dado que el trabajo se ha desarrollado en el contexto de un Trabajo Fin de Máster con dedicación parcial del autor, el coste principal corresponde a las horas de trabajo invertidas. La Tabla 6 recoge el desglose de costes.

Se ha estimado el coste de las horas de trabajo aplicando una tarifa de referencia de 25 USD/hora, alineada con el promedio de mercado para perfiles de ciencia de datos en Latinoamérica. El uso predominante de herramientas de código abierto ha permitido minimizar significativamente los costes de software, siendo Tableau Desktop la única herramienta con coste de licencia comercial. El acceso a la base de datos corporativa y la dirección académica no han supuesto costes adicionales directos para el proyecto.

Tipo de coste	Valor	Comentarios
Horas de trabajo en el proyecto	5,500 – 6,500 $	220-260 horas estimadas x Tarifa de referencia: 25 USD/h.
Equipo técnico	800 – 1,200 $	Ordenador portátil personal. Valor de mercado estimado (amortización parcial).
Software utilizado	480 $	Herramientas de código abierto: 0 USD. Tableau Desktop: ~480 USD/año (licencia Creator)
Estudios e informes	0$	Referencias de acceso abierto o vía biblioteca.

Materiales empleados	0$	Proyecto íntegramente digital.
Total estimado		6,780 – 8,180 $

Tabla 6: Desglose de costes del proyecto. Fuente: Elaboración propia.

4.5 Viabilidad

El proyecto es viable desde tres puntos de vista complementarios: técnico, económico y de sostenibilidad.

Desde el punto de vista de la viabilidad técnica, el stack tecnológico empleado (Python, scikit-learn, XGBoost) es un ecosistema maduro, ampliamente documentado y adoptado en la industria para problemas de clasificación con datos tabulares. Los datos necesarios para la construcción del modelo estaban disponibles en las fuentes corporativas de Nacional Seguros y no se necesitó de infraestructura de cómputo especializada, lo cual confirma la factibilidad técnica del enfoque seleccionado.

En términos económicos, el mayor costo del proyecto es el tiempo del analista. El retorno potencial se mide por la mejora de eficiencia de las campañas comerciales: si el modelo permite concentrar el esfuerzo de venta en un subconjunto reducido de clientes con mayor propensión, la reducción del coste por contacto improductivo puede justificar la inversión realizada. En la sección de resultados, se muestran los resultados específicos de esta capacidad de priorización. Además, tanto el *pipeline* para la extracción y transformación de datos como el *pipeline* analítico de modelado y evaluación están diseñados de forma modular, de manera que se pueden utilizar de nuevo para otros productos del portafolio, cambiando las variables explicativas y la variable objetivo. Esta característica amplifica el retorno de la inversión, ya que el coste de desarrollo se amortiza con varias iniciativas de venta cruzada o retención. El uso mayoritario de herramientas de código abierto también ha servido para reducir los costes de software.

En cuanto a la sostenibilidad, el *pipeline* desarrollado es reproducible y puede ejecutarse de forma periódica con nuevos cortes temporales, lo que permite actualizar las puntuaciones de propensión a medida que se vayan acumulando nuevos datos transaccionales. El entregable de marketing ha sido diseñado para ser utilizado operativamente con el equipo de ventas de Nacional Seguros, asegurando de esta manera su continuidad de uso más allá del ámbito académico del presente trabajo.

4.6 Resultados del proyecto

En el presente apartado se describen los resultados obtenidos en cada una de las etapas del *pipeline* analítico desarrollado, desde la validación del *dataset* hasta la generación del entregable comercial de *scoring*. Los resultados se presentan en orden cronológico de ejecución y de acuerdo con la metodología descrita en el apartado 4.2.

4.6.1 Construcción y validación del dataset

El *dataset* consolidado se construyó a partir de consultas SQL sobre la base de datos corporativa de Nacional Seguros, de acuerdo con el diseño descrito en el apartado 4.2.3. El resultado fue un *dataset* con 1,268,176 registros y 46 columnas, distribuidas en 41 cortes mensuales entre junio de 2020 y noviembre de 2025, lo que proporciona una base temporal de aproximadamente 5.4 años para la detección de patrones de venta cruzada.

La validación del *dataset* confirmó que no existen registros duplicados ni valores nulos en las variables objetivo. El análisis de la distribución del *target* mostró un desbalance extremo de clases: 425 casos positivos para el horizonte de 3 meses (tasa base de 0,034%) y 809 casos positivos para el horizonte de 6 meses (tasa base de 0,064%), lo que implica ratios de desbalance de aproximadamente 2,980:1 y 1,568:1 respectivamente. Este hallazgo condicionó de forma determinante la selección de métricas de evaluación y la configuración del modelo.

4.6.2 Análisis exploratorio de datos

El análisis exploratorio permitió describir el perfil demográfico de la cartera: edad media 43.9 años (desviación estándar 11.85), distribución por género 49.6% hombres, 47.5% mujeres y 2.9% sin información, estado civil predominantemente soltero (50.4%) seguido de casado (37.9%). La antigüedad media de los contratantes era de 62.7 meses.

En cuanto a la penetración de productos se pudo observar que el 49% de los registros presentaba al menos una póliza de Vida vigente en los últimos 12 meses, el 30% de Salud y el 24% de Automotores. La tasa del evento objetivo (cp_compro_6m) mostró variabilidad temporal, con un pico de 0.16% en marzo de 2025, aproximadamente 2.5 veces la tasa base promedio, lo cual sugiere posibles efectos de campañas comerciales o estacionalidad de la demanda.

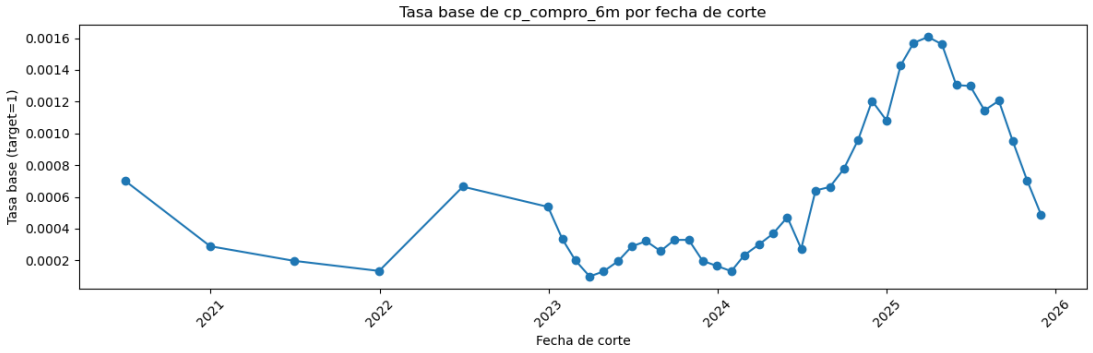

Figura 1. Tasa base de cp_compro_6m por fecha de corte. Fuente: elaboración propia (NB02).

4.6.3 Preparación para modelado

El análisis de calidad de datos identificó 835 registros (0,066%) con edades atípicas (valores negativos o superiores a 95 años), que fueron marcados mediante un flag binario (flag_edad_atipica) y sus valores originales sustituidos por nulo para su posterior imputación en

el *pipeline*. Adicionalmente, las variables de recencia contenían el valor centinela 99,999 para indicar ausencia de evento, el cual fue sustituido por valor nulo.

La etapa de preparación transformó el *dataset* de 46 columnas a 39, tras eliminar 6 columnas que representaban fugas de información temporal (fechas de contratación, número de pólizas y montos posteriores al corte del producto Casa Protegida) y 1 columna auxiliar de auditoría. De las 39 columnas resultantes, 4 corresponden a trazabilidad y control (IdContratante, FechaCorte, cp_compro_3m, cp_compro_6m) y 35 constituyen las variables candidatas para el modelado, distribuidas en: 3 demográficas, 3 *flags* de portafolio, 1 de antigüedad general, 11 de comportamiento en Salud, 5 de Vida, 8 de Automotores y 4 *flags* auxiliares generados durante la preparación. La descripción del preprocesamiento aplicado mediante ColumnTransformer se presentó en el apartado 4.2.4.

4.6.4 Modelo de propensión: entrenamiento y evaluación del modelo base

La partición de datos siguió un esquema temporal estricto para evitar fugas de información, aplicado de forma independiente para cada horizonte temporal. Para el horizonte de 6 meses, el conjunto de entrenamiento comprendió 882,559 registros con cortes hasta noviembre de 2024, se estableció un gap temporal de 6 meses (diciembre de 2024 a mayo de 2025) como periodo de separación, y el conjunto de *test* incluyó 96,997 registros con cortes de junio a agosto de 2025. Para el horizonte de 3 meses, el conjunto de entrenamiento comprendió 1,073,247 registros con cortes hasta mayo de 2025, con un gap de 3 meses y un conjunto de *test* de 97,932 registros con cortes de septiembre a noviembre de 2025. En ambos casos se definió un subconjunto de *test* de corte único (último corte disponible) para proporcionar una evaluación adicional bajo condiciones más cercanas al escenario operativo real.

El modelo base XGBoost se configuró con los parámetros descritos en el apartado 4.2.4. Tabla 1: Configuración del modelo base XGBoost para ambos horizontes temporales. Fuente: Elaboración propia. (Tabla 1). La Tabla 7 resume el rendimiento del modelo base en los cuatro conjuntos de evaluación.

Métrica	3m — test	3m — test_1_corte	6m — test	6m — test_1_corte
Registros	97932	32694	96997	32344
Positivos	52	16	118	39
Tasa base	0.0531%	0.0489%	0.1217%	0.1206%
PR-AUC	0.001006	0.001247	0.00238	0.003282
ROC-AUC	0.636385	0.633751	0.661941	0.697239

Tabla 7: Rendimiento del modelo base XGBoost por horizonte temporal y conjunto de evaluación. Fuente: elaboración propia.

Si bien los valores absolutos de PR-AUC resultan bajos en comparación con problemas de clasificación convencionales, deben interpretarse en relación con las tasas base indicadas: para el horizonte de 6 meses, el PR-AUC supera en aproximadamente el doble a la tasa base (0.0024 frente a 0.0012), lo que indica capacidad discriminativa útil para la priorización comercial. El

modelo de 6 meses muestra un rendimiento superior consistente en ambas métricas, lo cual resulta coherente con la mayor disponibilidad de ejemplos positivos en ese horizonte (118 frente a 52 en test). El rendimiento sobre el *test* de corte único es superior al del *test* completo en el horizonte de 6 meses (ROC-AUC de 0.6972 frente a 0.6619), lo que sugiere que el modelo preserva su capacidad discriminativa en el escenario más cercano al uso operativo.

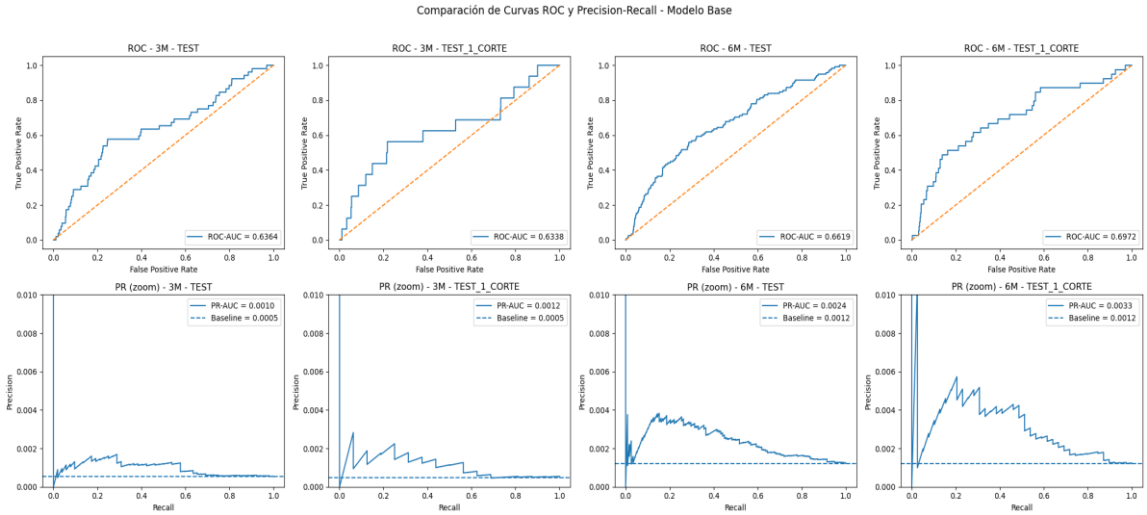

Ilustración 4: Curvas ROC y Precision-Recall del modelo base XGBoost, por horizonte temporal (3 y 6 meses) y conjunto de prueba (test completo y test de corte único). Fuente: elaboración propia.

La Ilustración 4 representa visualmente el rendimiento del modelo base XGBoost reportado en la Tabla 7. La fila superior muestra las curvas ROC para las cuatro combinaciones de horizonte temporal y conjunto de prueba, con el valor de ROC-AUC indicado en cada panel. La fila inferior muestra las curvas Precision-Recall con ampliación en los rangos bajos de precisión, consistentes con el nivel de desbalance del problema; en cada panel se reporta el PR-AUC y la línea base horizontal correspondiente a la tasa de positivos del conjunto evaluado.

Se observa que el modelo base alcanza una separación clara respecto a la diagonal de azar en las curvas ROC y un PR-AUC varias veces superior a la línea base en todos los conjuntos, lo que confirma una señal predictiva aprovechable incluso antes de la optimización de hiperparámetros. Este rendimiento establece la línea de referencia frente a la cual se evaluará, en apartados posteriores, la mejora obtenida con el modelo optimizado.

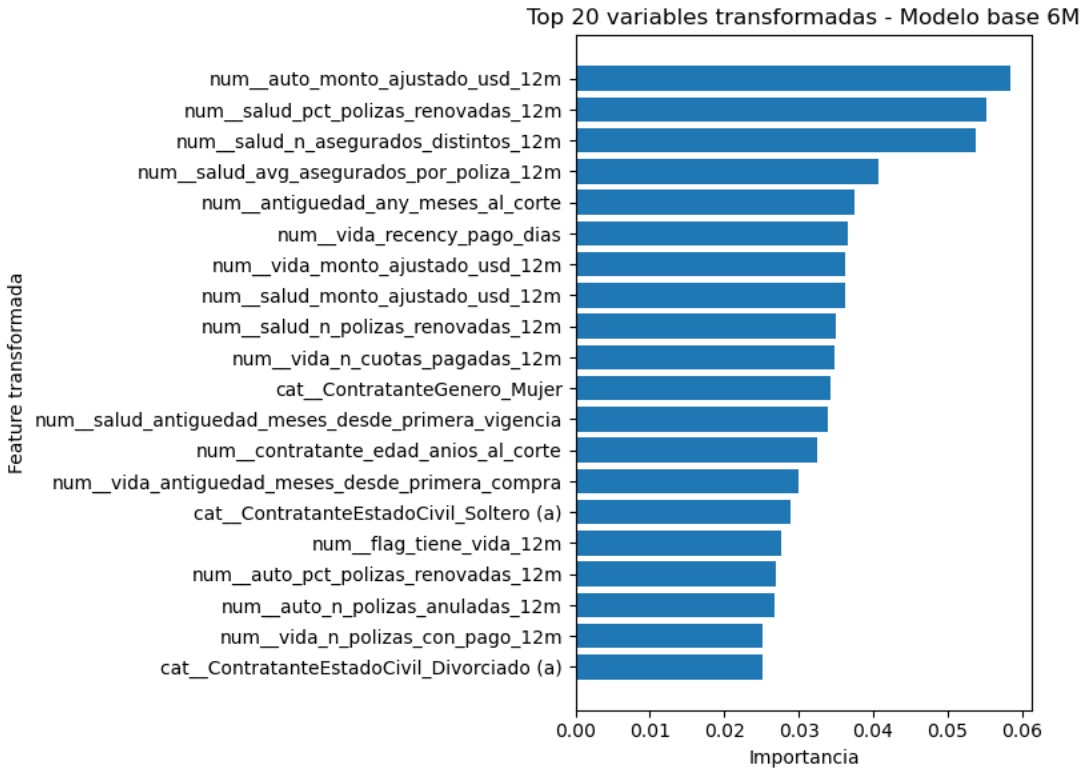

Ilustración 5: Top 20 variables por importancia (ganancia) del modelo base XGBoost — horizonte de 6 meses. Fuente: elaboración propia.

La Ilustración 5 muestra las 20 variables con mayor importancia por ganancia del modelo base XGBoost para el horizonte de 6 meses. Los tres predictores dominantes son el monto ajustado en dólares asociado al ramo de Automotores en los últimos doce meses (num__auto_monto_ajustado_usd_12m), el porcentaje de pólizas renovadas en el ramo de Salud (num__salud_pct_polizas_renovadas_12m) y el número de asegurados distintos cubiertos bajo pólizas de Salud (num__salud_n_asegurados_distintos_12m). A estos se suman, en posiciones altas del ranking, variables monetarias de los tres ramos (Automotores, Salud y Vida), métricas de recencia de pagos en Vida y variables de antigüedad del contratante. Las variables demográficas (edad, género y estado civil) aparecen en el top 20 con pesos menores, lo que indica que la señal predictiva del modelo base se concentra principalmente en el comportamiento transaccional del contratante en los ramos ya contratados, antes que en sus atributos demográficos.

4.6.5 Optimización de hiperparámetros

La optimización se llevó a cabo mediante GridSearchCV con las 128 combinaciones de hiperparámetros definidas en la Tabla 2 (apartado 4.2.4), evaluadas con validación cruzada cronológica de 3 *folds* mediante PredefinedSplit y PR-AUC (average_precision) como métrica de optimización.

La Tabla 8 presenta los hiperparámetros óptimos seleccionados para cada horizonte.

Hiperparámetro	6 meses	3 meses
n_estimators	300	300
learning_rate	0.03	0.03
max_depth	5	3
min_child_weight	1	1
subsample	0.8	1
colsample_bytree	0.8	1
reg_lambda	1.0	1

Tabla 8: Hiperparámetros óptimos por horizonte temporal. Fuente: elaboración propia.

La Tabla 9 compara el rendimiento del modelo optimizado frente al modelo base en los cuatro conjuntos de evaluación.

Test set	Métrica	Base	Nuevo tuned	Δ vs base
3m test	PR-AUC	0.001006	**0.001857**	+84.6%
3m test	ROC-AUC	0.636385	**0.743091**	+16.8%
3m 1_corte	PR-AUC	0.001247	**0.003891**	+211.9%
3m 1_corte	ROC-AUC	0.633751	**0.752920**	+18.8%
6m test	PR-AUC	0.002380	**0.003327**	+39.8%
6m test	ROC-AUC	0.661941	**0.726734**	+9.8%
6m 1_corte	PR-AUC	0.003282	**0.004408**	+34.3%
6m 1_corte	ROC-AUC	0.697239	**0.770062**	+10.4%

Tabla 9: Comparación del rendimiento entre modelo base y modelo optimizado en los cuatro conjuntos de evaluación. Fuente: elaboración propia.

La mejora del horizonte de 3 meses fue consistentemente más pronunciada (+84.6% en PR-AUC y +16.8% en ROC-AUC sobre test completo) que la del de 6 meses (+39.8% y +9.8%), resultado coherente con que el modelo base de 3 meses partiera de un rendimiento inferior y se beneficiara proporcionalmente más del ajuste de hiperparámetros. En ambos horizontes, la mejora se mantiene tanto en el *test* completo como en el *test* de corte único, donde el modelo de 6 meses alcanza un ROC-AUC de 0.7701 y el de 3 meses un ROC-AUC de 0.7529, confirmando la ganancia discriminativa en el escenario más cercano al uso operativo. Cabe destacar que la corrección de la validación cruzada — incorporando un periodo de gap equivalente al horizonte del *target* entre los *folds* de entrenamiento y validación — fue determinante para seleccionar hiperparámetros que generalizan mejor sobre datos futuros no vistos.

Los pipelines finales (preprocesamiento + modelo optimizado) se serializaron mediante joblib para su reutilización directa en *scoring*.

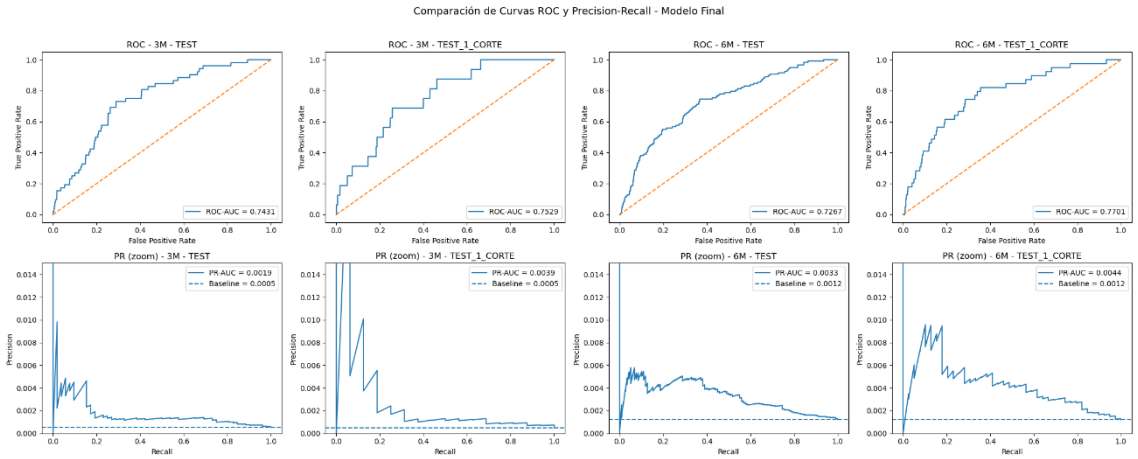

Ilustración 6: Curvas ROC y Precision-Recall del modelo final optimizado, por horizonte temporal (3 y 6 meses) y conjunto de prueba (test completo y test de corte único). Fuente: elaboración propia.

La Ilustración 6 presenta el rendimiento del modelo final optimizado en las cuatro combinaciones de horizonte temporal y conjunto de prueba, con la misma estructura visual utilizada en la Ilustración 2. La fila superior muestra las curvas ROC con su respectivo ROC-AUC, y la fila inferior las curvas Precision-Recall con ampliación en el rango relevante para este problema desbalanceado, indicando en cada panel el PR-AUC alcanzado y la línea base correspondiente.

El modelo optimizado alcanza valores de ROC-AUC comprendidos entre 0.7267 y 0.7701 en los cuatro escenarios, confirmando una capacidad discriminativa moderada y estable. El PR-AUC se sitúa entre 0.0019 y 0.0044, valores reducidos en términos absolutos, pero entre 3.8 y 7.8 veces superiores a la línea base de cada conjunto. El mejor rendimiento conjunto se observa en el horizonte de 6 meses evaluado sobre el *test* de corte único (ROC-AUC 0.7701; PR-AUC 0.0044), lo que respalda la decisión de emplear este horizonte para la generación del entregable comercial analizado en los apartados posteriores.

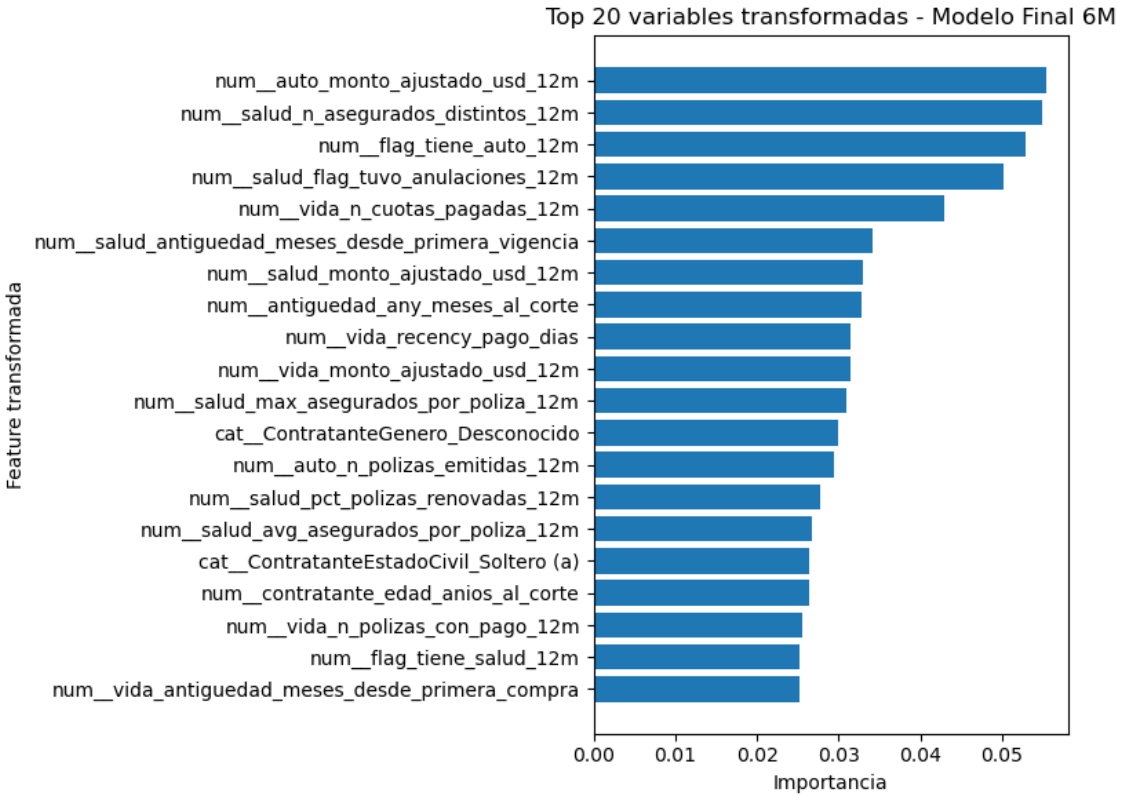

Ilustración 7: Top 20 variables por importancia (ganancia) del modelo final optimizado XGBoost — horizonte de 6 meses. Fuente: elaboración propia.

La Ilustración 7 presenta las 20 variables con mayor importancia por ganancia del modelo final optimizado para el horizonte de 6 meses. La variable líder se mantiene respecto al modelo base (num__auto_monto_ajustado_usd_12m), lo que confirma la estabilidad del principal predictor asociado a la intensidad del vínculo comercial en el ramo de Automotores. Sin embargo, la optimización de hiperparámetros reordena el ranking y eleva a posiciones superiores indicadores de tenencia y comportamiento categórico, como el *flag* de tenencia en Automotores (num__flag_tiene_auto_12m), el *flag* de anulaciones previas en Salud (num__salud_flag_tuvo_anulaciones_12m) y el número de cuotas pagadas en Vida durante los últimos doce meses (num__vida_n_cuotas_pagadas_12m). Las variables de antigüedad y demográficas se conservan en el top 20 pero con pesos inferiores. Esta redistribución refina la estructura explicativa del modelo sin alterar su núcleo predictivo, priorizando señales de composición de portafolio y de comportamiento histórico por encima de métricas porcentuales agregadas. El detalle numérico completo de las 20 variables se reporta en la Tabla 14 del Anexo B.

4.6.6 Evaluación Top-K y capacidad de priorización

Para cuantificar la utilidad del modelo como herramienta de priorización comercial, se realizó un análisis Top-K sobre los conjuntos de *test* de corte único, que representan el escenario más

próximo al uso operativo. La Tabla 10 presenta los resultados para el horizonte de 6 meses (32,344 clientes, 39 positivos observados, tasa base 0.1206%).

Top K	Seleccionados	TP	Recall	*Lift*
1%	323	1	2.6%	2.57
3%	970	7	17.9%	5.98
5%	1,617	8	20.5%	4.10
10%	3,234	16	41.0%	4.10
20%	6,468	24	61.5%	3.08

Tabla 10: Análisis Top-K del modelo optimizado — Horizonte 6 meses, test de corte único. Fuente: elaboración propia.

El Top 1% captura un único positivo (*Lift* de 2.57×), lo que refleja la dificultad inherente de concentrar aciertos en un segmento tan reducido (323 clientes) con solo 39 eventos observados. A partir del Top 3%, el modelo demuestra una capacidad de concentración significativa: con un *Lift* de 5.98×, la probabilidad de encontrar un caso positivo dentro de los 970 clientes mejor puntuados es casi seis veces superior a la selección aleatoria. En el Top 10% (3,234 clientes) se capturan 16 de los 39 positivos con un Recall del 41.0%, y en el Top 20% se alcanza un Recall del 61.5%, lo que indica que casi dos tercios de los compradores potenciales se concentran en el quintil de mayor puntuación.

La Tabla 11 presenta los resultados equivalentes para el horizonte de 3 meses (32,694 clientes, 16 positivos observados, tasa base 0.0489%)

Top K	Seleccionados	TP	Recall	*Lift*
1%	326	2	12.5%	12.54
3%	980	3	18.8%	6.26
5%	1,634	3	18.8%	3.75
10%	3,269	5	31.3%	3.13
20%	6,538	8	50.0%	2.50

Tabla 11: Análisis Top-K del modelo optimizado — Horizonte 3 meses, test de corte único. Fuente: elaboración propia.

El horizonte de 3 meses presenta un *Lift* destacable en el Top 1% (12.54×), sustentado en 2 casos positivos capturados sobre 16 posibles. Si bien el reducido número de eventos limita la fiabilidad estadística de esta cifra, el resultado sugiere que el modelo concentra la señal de propensión de forma efectiva en el extremo superior del ranking. En el Top 20%, el Recall alcanza el 50.0%, cifra comparable a la obtenida para 6 meses (61.5%). Ambos horizontes confirman que la concentración de esfuerzo comercial en los segmentos de mayor puntuación permite multiplicar la eficiencia de contacto entre 2.5× y 6.0× respecto a campañas no dirigidas.

En la ilustración 8 presenta la matriz de confusión correspondiente al modelo optimizado de 6 meses sobre el *test* de corte único, utilizando como umbral de clasificación el Top 10% (3,235 clientes seleccionados). El modelo identifica correctamente 16 de los 39 casos positivos (Recall

del 41.0%), a costa de 3,219 falsos positivos. Si bien la tasa de falsos positivos puede parecer elevada en términos absolutos, en el contexto de una campaña comercial de *cross-sell* esta relación resulta operativamente aceptable: contactar a 3,235 clientes para capturar 16 compradores potenciales implica un ratio de contacto de 202:1, frente a los 829:1 que supondría una campaña no dirigida sobre los 32,344 clientes (ratio derivado de la tasa base de 0.12%). Dicho de otro modo, el esfuerzo comercial por conversión se reduce a menos de un cuarto.

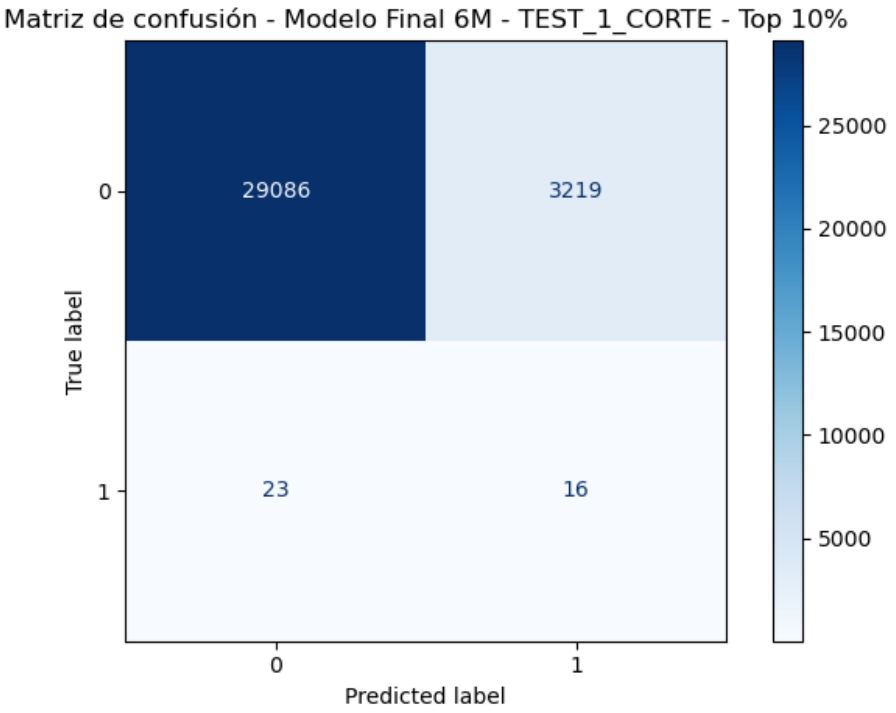

Ilustración 8: Matriz de confusión — Modelo optimizado 6 meses, test de corte único (Top 10%). Fuente: elaboración propia.

4.6.7 Scoring y entregable comercial

El *pipeline* final optimizado se aplicó al corte más reciente disponible, generando puntuaciones de propensión para 32,344 clientes activos. Las puntuaciones se distribuyeron entre un mínimo de 0,001 y un máximo de 0,952, con una media de 0,128. A partir de estas puntuaciones se definieron seis bandas de priorización basadas en percentiles de ranking.

Banda	Clientes	Rango de score	Prioridad
Top 0.5%	161	0,836 – 0,952	Crítica — contacto ejecutivo inmediato
Top 0.5 – 1%	162	0,781 – 0,836	Muy alta
Top 1 – 2%	323	0,701 – 0,781	Alta — campaña dirigida
Top 2 – 5%	971	0,537 – 0,701	Media — campaña digital ampliada
Top 5 – 10%	1,617	0,361 – 0,537	Media-baja — campaña digital amplia
Resto	29,110	< 0,361	No priorizados en primera ola

Tabla 12: Bandas de priorización por score de propensión. Fuente: elaboración propia.

El archivo conserva siete hojas: Resumen (indicadores ejecutivos del *scoring*), ResumenBandas (métricas agregadas por banda), ResumenTopK (análisis de Precision, Recall y *Lift* por percentil), Top_0_5 (listado nominal de los 161 clientes de máxima prioridad), Top_1 (323 clientes), Top_5 (1,617 clientes) y UniversoScoreado (los 32,344 clientes con 20 columnas de información que incluyen score, ranking, banda, segmento de prioridad, recomendación de acción, perfil de portafolio, datos demográficos, flags de producto y montos de prima). Este formato permite al equipo comercial de Nacional Seguros segmentar y priorizar las acciones de *cross-sell* de forma inmediata y autónoma, sin requerir conocimientos técnicos para su explotación.

4.6.8 Visualización analítica

Como complemento al entregable comercial descrito en el apartado anterior, se diseñó un cuadro de mando interactivo en Tableau Desktop que permite explorar los resultados del *scoring* desde tres perspectivas complementarias: rendimiento del modelo, segmentación de la cartera y acción comercial.

Para efectos de la entrega académica, el *dashboard* se publicó en Tableau Public y se encuentra accesible en la siguiente dirección:

https://public.tableau.com/shared/KCKHTCCDK?:display_count=n&:origin=viz_share_link.

Dentro del contexto operativo de Nacional Seguros, el cuadro de mando está pensado para ser publicado en el servidor on-premise de Tableau de la compañía, donde podrá integrarse con los flujos de trabajo del equipo comercial y actualizarse con cada nuevo ciclo de *scoring*.

El cuadro de mando opera sobre los 32,344 registros del universo scoreado correspondientes al corte de agosto de 2025 y se estructura en tres pestañas temáticas. Las tres pestañas comparten un panel superior de indicadores clave que presenta de forma permanente cinco métricas de referencia: el número de clientes scoreados (32,344), el score máximo observado (0,952), el score promedio (0,128), el número de positivos observados en el horizonte de seis meses (39) y el *Lift* acumulado en el Top 2% de la cartera (6.4×).

Es necesario señalar que la estructura de bandas utilizada en el *dashboard* difiere de la empleada durante la evaluación del modelo en los *notebooks*. En la fase de evaluación (apartado 4.6.6, Tabla 10), se utilizaron percentiles Top 1%, 3%, 5%, 10% y 20% sobre el conjunto de *test*, con el propósito de medir la capacidad discriminativa del modelo en condiciones controladas. En cambio, el *dashboard* adopta los percentiles Top 0.5%, 1%, 2%, 5% y 10% definidos en las bandas de priorización del entregable comercial (apartado 4.6.7, Tabla 12), calculados sobre el universo completo de 32,344 clientes. Esta diferencia es intencional: las bandas de evaluación responden a una lógica de validación estadística del modelo, mientras que las bandas del *dashboard* responden a una lógica operativa orientada a la segmentación comercial, donde los cortes más finos en la parte superior de la distribución (Top 0.5% y Top 1%) resultan más útiles para la priorización de acciones de contacto.

4.6.8.1 Pestaña de rendimiento del modelo

La primera pestaña presenta la distribución del *scoring* y las métricas de rendimiento por banda. El histograma de distribución de contratantes por score permite visualizar la concentración marcada en el rango inferior (0.00 a 0.10) y la cola derecha progresivamente más delgada, consistente con la baja prevalencia del evento objetivo (tasa base del 0.12%). El gráfico de doble eje de Recall y *Lift* por Top K y la tabla de métricas por banda complementan la información presentada en las Tablas 10 y 12 de los apartados anteriores, permitiendo al usuario explorar interactivamente la información respecto al performance del modelo.

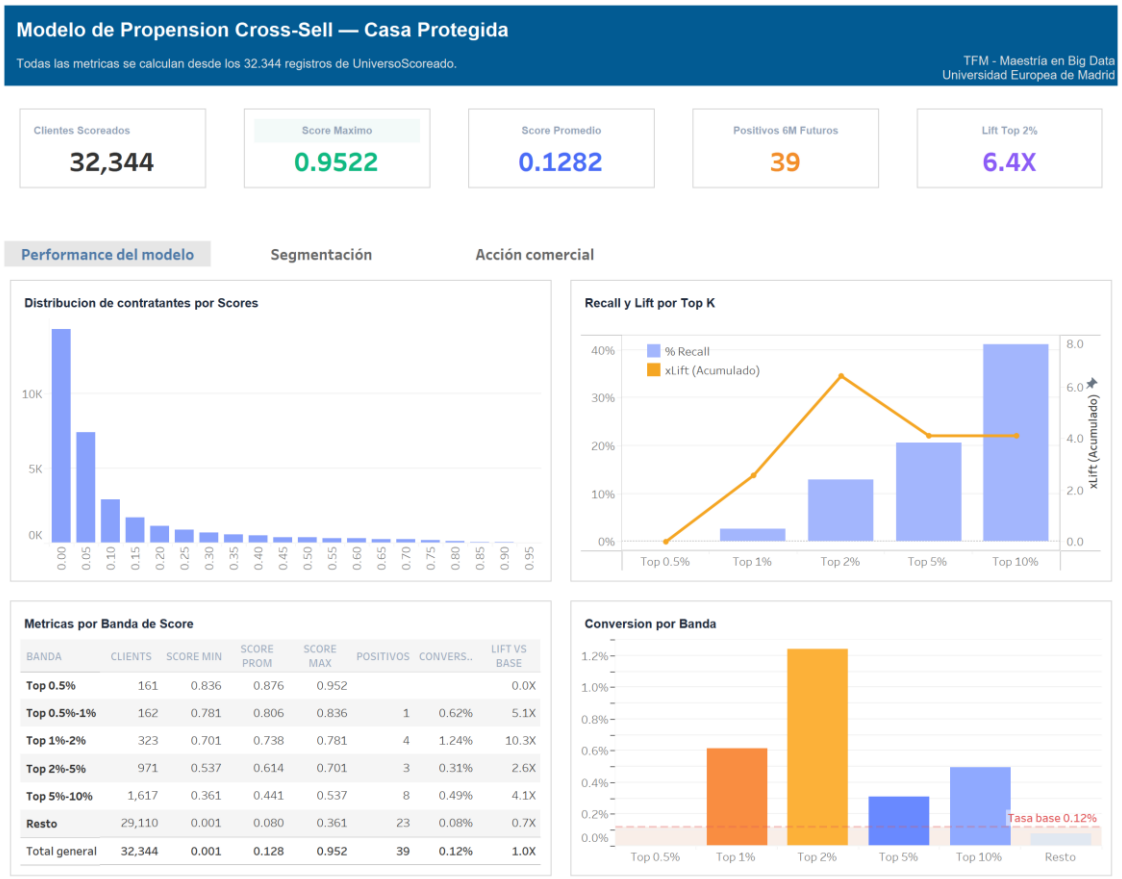

Ilustración 9: Vista de rendimiento del modelo — Dashboard interactivo en Tableau. Fuente: elaboración propia.

4.6.8.2 Pestaña de segmentación

La segunda pestaña caracteriza el universo calificado en tres dimensiones: portafolio de productos, perfil demográfico y comportamiento de conversión.

El gráfico de conversión por perfil de portafolio revela que los contratantes con la combinación Salud/Auto presentan la mayor tasa de conversión observada al producto Casa Protegida, seguidos por los contratantes con solo Salud y los contratantes con solo Vida. Este hallazgo sugiere que la tenencia simultánea de productos en más de una línea de negocio, particularmente cuando incluye Salud y Automotores, se asocia con una mayor propensión a la

contratación de Casa Protegida, resultado coherente con las tres variables de mayor importancia del modelo: auto_monto_ajustado_usd_12m, flag_tiene_auto_12m y salud_n_asegurados_distintos_12m (Tabla 14, Anexo B).

El gráfico de doble eje de conversión y score promedio por antigüedad muestra una tendencia creciente en ambas métricas a medida que aumenta el tiempo transcurrido desde la primera contratación. Los contratantes con antigüedad entre 60 y 120 meses presentan los valores más altos tanto de conversión como de score promedio, lo que indica que la relación de largo plazo con la aseguradora constituye un factor relevante para la venta cruzada de Casa Protegida, consistente con la presencia de antiguedad_any_meses_al_corte entre las ocho variables más importantes del modelo (Tabla 14, Anexo B).

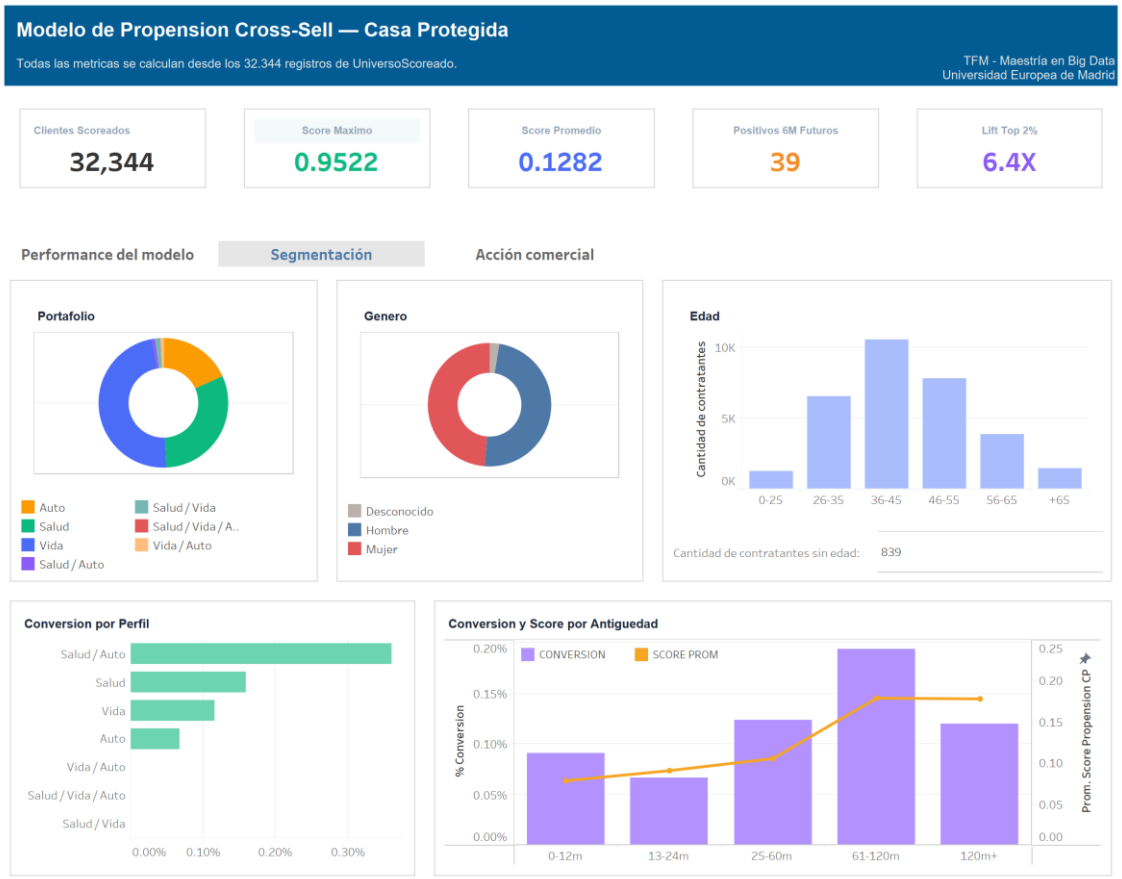

Ilustración 10: Vista de segmentación — Dashboard interactivo en Tableau. Fuente: elaboración propia.

4.6.8.3 Pestaña de acción comercial

La tercera pestaña está orientada a la toma de decisiones operativas. Su componente principal es la Matriz de Acción Comercial, que sintetiza en una única tabla la información de la Tabla 12 del apartado 4.6.7 junto con las recomendaciones de intervención por banda, proporcionando al equipo de marketing una guía directa para la asignación de recursos.

El gráfico de prima promedio en USD por línea de negocio y banda aporta una dimensión que no se aborda en las secciones anteriores: el valor monetario asociado a cada segmento de priorización. Se observa que los contratantes de mayor propensión tienden a concentrar primas más elevadas, particularmente en la línea de Salud. Este hallazgo refuerza el argumento económico para la focalización comercial en las bandas superiores: estos segmentos no solo agrupan a los clientes con mayor probabilidad de compra, sino también a aquellos que ya representan un mayor volumen de negocio para la compañía. El gráfico de volumen por tipo de intervención complementa esta vista mostrando que los esfuerzos comerciales priorizados se concentran en 3,234 contratantes (10.1% de la cartera), mientras que el 89.9% restante queda fuera de la primera ola de acción.

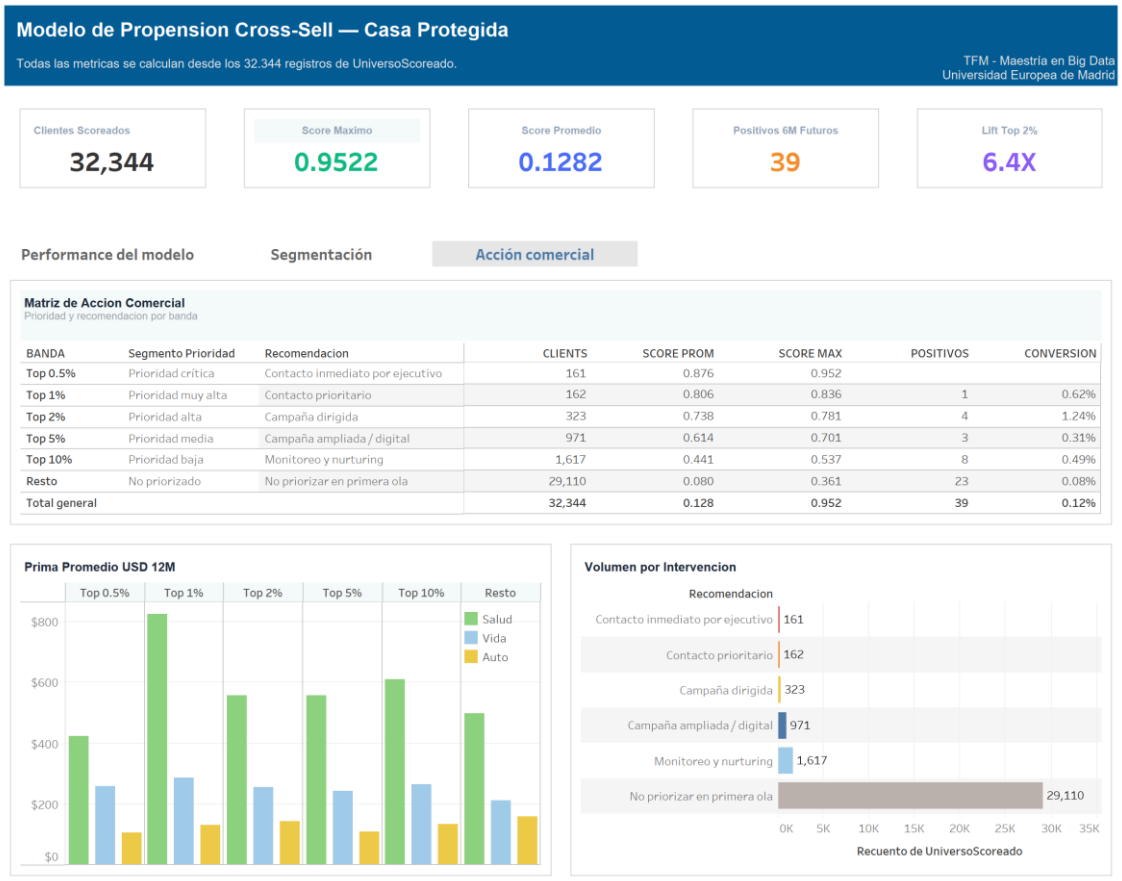

Ilustración 11: Vista de acción comercial — Dashboard interactivo en Tableau. Fuente: elaboración propia.

4.6.9 Estimación del impacto económico del sistema de priorización

Con el fin de proyectar el valor que puede aportar el sistema de valoración y priorización desarrollado a Nacional Seguros, se presenta a continuación una estimación comparativa del impacto económico de dos escenarios de gestión comercial del producto Casa Protegida: uno basado en el modelo de propensión (escenario priorizado) y otro sin priorización alguna (escenario masivo). La estimación se hace sobre el universo scoreado de 32,344 clientes

correspondiente al corte de agosto de 2025 y se toma como referencia una prima promedio de 120 USD por póliza contratada.

Los parámetros operativos utilizados en ambos escenarios se recogen en la Tabla 13.

Parámetro	Valor	Observación
Universo total	32,344 clientes	Corte de evaluación operativo (agosto 2025)
Prima promedio Casa Protegida	120 USD	Valor de referencia por póliza anual
Coste de envío WhatsApp	0.07 USD por mensaje	Canal de contacto inicial
Coste de llamada telefónica	5 Bs (~0.72 USD)	Gestión comercial directa
Tasa de respuesta al WhatsApp	5%	Supuesto operativo
Positivos observados en el universo	39	Compradores en ventana de 6 meses
Positivos en el Top 20% (modelo)	24	Recall@20% = 61.5%; *Lift* = 3.08×
Positivos esperados en un 20% aleatorio	~8	Proporcional a la tasa base (0.12%)

Tabla 13:Supuestos operativos para la estimación de impacto económico. Fuente: elaboración propia.

En el escenario priorizado, el modelo selecciona el 20% de la cartera con mayor score de propensión (6,468 clientes). De estos, los 161 clientes del Top 0.5% reciben contacto telefónico directo por parte de un ejecutivo comercial, conforme a la recomendación de acción definida en el apartado 4.6.7. Los 6,307 clientes restantes del grupo priorizado reciben un mensaje de WhatsApp; de ellos, se asume que un 5% manifiesta interés (315 clientes), a quienes se contacta telefónicamente.

En el escenario masivo (sin modelo), se envía el mensaje de WhatsApp a los 32,344 clientes del universo completo. El 5% que manifiesta interés (1,617 clientes) recibe una llamada telefónica. Este escenario no contempla llamadas directas, dado que no se dispone de un criterio de priorización para identificar a los clientes de mayor potencial.

La Tabla 14 compara los costes operativos de ambos escenarios.

Concepto	Escenario priorizado	Escenario masivo

Mensajes de WhatsApp	6,307 × 0.07 = 441 USD	32,344 × 0.07 = 2,264 USD
Llamadas telefónicas	(315 + 161) × 5 Bs = 2,380 Bs (~342 USD)	1,617 × 5 Bs = 8,085 Bs (~1,162 USD)
Coste total por campaña	~783 USD	~3,426 USD
Ahorro operativo	2,643 USD (77%)	—

Tabla 14: Comparación de costes operativos por campaña semestral. Fuente: elaboración propia.

El modelo reduce el coste de la campaña en un 77% y el número de llamadas telefónicas de 1,617 a 476, lo que implica una liberación significativa de la capacidad del equipo comercial.

Para estimar el ingreso incremental, se compara la captación de compradores potenciales en cada escenario. El Top 20% del modelo concentra 24 de los 39 compradores observados en la ventana de seis meses (Recall@20% = 61.5%), mientras que una selección aleatoria del mismo tamaño captaría proporcionalmente unos 8 compradores. La diferencia de 16 compradores constituye la ventaja de captación atribuible al modelo. Asumiendo que la totalidad de los compradores identificados en el grupo priorizado se concreta mediante la gestión comercial, el ingreso incremental por campaña asciende a 16 × 120 = 1,920 USD.

Adicionalmente, el segmento priorizado (Top 5%) incluye 1,609 clientes que, pese a no haber adquirido el producto durante el periodo de observación, presentan los perfiles de propensión más elevados según el modelo. Este grupo representa una reserva de demanda cuya conversión, incluso parcial, incrementaría el retorno estimado. A modo ilustrativo, la conversión de un 1% adicional de este segmento representaría 16 pólizas adicionales (1,920 USD).

La Tabla 15 consolida la estimación de valor anualizado del sistema, asumiendo dos campañas por año.

Concepto	Por campaña (USD)	Anualizado (USD)
Ahorro operativo	2,643	5,286
Ingreso incremental (captación diferencial)	1,920	3,840
Ingreso potencial (demanda latente, 1%)	1,920	3,840
Valor total estimado	6,483	12,966
Coste del proyecto (Tabla 6)	—	6,780 – 8,180
Retorno sobre la inversión (ROI) año 1	—	59% – 91%

Periodo de recuperación estimado	—	8 – 10 meses

Tabla 15: Estimación de valor anual del sistema de priorización. Fuente: elaboración propia.

Hay que aclarar que los valores de ingreso incremental y de demanda están basados en supuestos que no han sido comprobados a través de una campaña real. La tasa de respuesta del canal de WhatsApp del 5%, la captación total de los compradores identificados y la conversión de la demanda latente del 1% son estimaciones de escenario favorable que requerirían una prueba piloto controlada para su confirmación.

Por último, el rendimiento estimado es únicamente para el producto Casa Protegida. Como el *pipeline* analítico ha sido desarrollado de forma modular (apartado 4.5), su aplicación a otros productos del portafolio de Nacional Seguros (Automotores, Vida, Salud) incrementaría el valor generado, con un coste marginal reducido, gracias a la reutilización de la infraestructura de datos, el código de modelado y la estructura del entregable comercial.

Capítulo 5. DISCUSIÓN

5.1 Interpretación y discusión de los resultados

El modelo XGBoost desarrollado muestra una capacidad discriminativa coherente con las condiciones de desbalance extremo del problema. Con una tasa base de contratación de Casa Protegida del 0,121% en un horizonte de 6 meses y del 0,049% en 3 meses (evaluadas sobre el *test* de corte único), los ratios de clases positivas frente a negativas ascienden a 1:829 y 1:2,045 respectivamente. En este contexto, las métricas globales deben interpretarse con cautela: el ROC-AUC del modelo optimizado de 6 meses alcanzó 0.7701 en el *test* de corte único (escenario operativo), lo que indica una capacidad de ordenamiento claramente superior a la aleatoria, si bien insuficiente para establecer umbrales binarios de clasificación fiables. Es la evaluación Top-K la que refleja de forma más directa el valor comercial del modelo.

Los resultados Top-K del modelo optimizado muestran que el segmento del 3% de clientes con mayor propensión asignada concentra una proporción de compradores efectivos casi seis veces superior a la que se obtendría mediante una selección aleatoria equivalente (*Lift* de 5.98× para el horizonte de 6 meses en el *test* de corte único). En el Top 5% el *Lift* se mantiene en 4.10× y en el Top 20% el modelo captura el 61.5% de los compradores potenciales (Recall). El Top 1% captura un único positivo entre 323 clientes (*Lift* de 2.57×), lo que indica que la capacidad discriminativa del modelo, aunque superior a la aleatoria en todos los percentiles, alcanza su máxima concentración en la banda del 3 al 10%. Este comportamiento es esperable en problemas con 39 eventos positivos en el conjunto de evaluación, donde la dispersión estadística en los percentiles extremos es inherentemente elevada.

La comparación entre horizontes temporales aporta hallazgos relevantes. El modelo de 3 meses exhibió una mejora absoluta mayor con la optimización de hiperparámetros (+10.7 puntos porcentuales en ROC-AUC sobre test completo y +11.9 sobre test de corte único) que el de 6 meses (+6.5 y +7.3 respectivamente). Esto resulta coherente con el hecho de que el modelo base de 3 meses partía de un rendimiento inferior (ROC-AUC de 0,636 frente a 0,662 en test) por contar con menos ejemplos positivos en evaluación (52 frente a 118). En cuanto a los hiperparámetros seleccionados, el modelo de 6 meses adoptó una configuración más compleja y regularizada (max_depth = 5, colsample_bytree = 0.8 y subsample = 0.8), mientras que el de 3 meses optó por árboles superficiales sin submuestreo (max_depth = 3, colsample_bytree = 1.0, subsample = 1.0). Esto sugiere que el mayor volumen de positivos del horizonte de 6 meses permitió al algoritmo explorar particiones más profundas, compensando la complejidad adicional mediante submuestreo de filas y columnas.

Un hallazgo relevante del *tuning* es la mejora consistente tanto en ROC-AUC como en PR-AUC en todos los conjuntos de evaluación. En el modelo optimizado de 6 meses evaluado sobre el test de corte único, el ROC-AUC aumentó 7.3 puntos porcentuales (de 0,697 a 0,770) y el PR-AUC creció un 34.3% (de 0.003282 a 0.004408). Esta mejora se materializó de forma especialmente visible en la evaluación Top-K: el segmento Top 3%, que en el modelo base presentaba un *Lift* de 0.85× (inferior a la selección aleatoria, con un único positivo capturado

entre 970 clientes), alcanzó un *Lift* de 5.98× en el modelo optimizado (7 positivos). Del mismo modo, el Recall del Top 10% pasó del 33.3% al 41.0% y el del Top 20% del 51.3% al 61.5%, confirmando que la optimización mejoró tanto el ordenamiento global como la concentración de positivos en los segmentos de mayor interés comercial.

El entregable comercial generado — 32,344 clientes segmentados en seis bandas de prioridad para el corte de agosto de 2025 — constituye el resultado operativo central del proyecto. El análisis Top-K sobre datos no vistos durante el entrenamiento avala la utilidad del modelo para guiar campañas de *cross-sell* de forma más eficiente que la selección no dirigida, reduciendo el ratio de contacto por conversión de 829:1 (selección aleatoria) a 139:1 (Top 3%) o 202:1 (Top 5%).

5.2 Limitaciones del estudio y adaptaciones metodológicas

El desbalance extremo de clases representa la limitación más relevante del estudio. La escasa frecuencia del evento genera alta dispersión estadística en un conjunto con solo 39 eventos positivos, lo que resta estabilidad a las estimaciones de probabilidad en los percentiles extremos, como se refleja en el hecho de que el Top 1% captura solo 1 positivo entre 323 clientes evaluados (*Lift* de 2.57×), a pesar de la configuración del parámetro scale_pos_weight y la adopción de métricas orientadas a priorización (PR-AUC, Top-K, *Lift*). De la misma manera, los valores de PR-AUC absolutos no son directamente comparables con estudios llevados a cabo en contextos de mayor prevalencia.

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

El presente trabajo ha desarrollado un sistema de *scoring* de propensión al *cross-sell* del producto Casa Protegida en Nacional Seguros, basado en un modelo supervisado de clasificación binaria con XGBoost. El sistema cubre el ciclo completo desde la construcción del *dataset* analítico hasta la generación de un entregable operativo y su visualización analítica, estructurado bajo los principios de la metodología CRISP-DM.

A modo de síntesis visual, la Ilustración 12 resume los seis hitos principales del *pipeline* desarrollado, organizados en dos fases: la fase de construcción —que abarca desde la consolidación del *dataset* hasta el modelado— y la fase de resultados —que comprende la evaluación del modelo seleccionado, la generación del entregable comercial y la visualización analítica en Tableau.

Ilustración 12: Síntesis del proyecto — hitos y resultados principales del pipeline analítico desarrollado. Fuente: elaboración propia.

Los objetivos específicos planteados se han cumplido de la siguiente manera. El histórico de ventas de contratantes naturales fue consolidado en un *dataset* reproducible de 1,268,176 registros y 46 columnas, construido mediante cortes temporales mensuales sobre la base de datos corporativa. El conjunto de datos fue preparado y transformado mediante un *pipeline* de scikit-learn con ColumnTransformer, resultando en 35 variables candidatas para el modelado. El modelo supervisado de propensión fue desarrollado y optimizado mediante GridSearchCV con 128 combinaciones de hiperparámetros y validación cruzada cronológica, para ambos horizontes temporales evaluados (3 y 6 meses). El modelo optimizado de 6 meses alcanzó un ROC-AUC de 0.7701 en el *test* de corte único y un *Lift* de 4.10× en el Top 5%, confirmando su

capacidad de concentrar compradores potenciales en los segmentos priorizados. El modelo de 3 meses obtuvo un ROC-AUC de 0.7529 en el *test* de corte único, con mejoras absolutas mayores respecto al modelo base (+11.9 puntos porcentuales en ROC-AUC frente a +7.3 del modelo de 6 meses), consistentes con el mayor margen de mejora derivado de un rendimiento base inferior. El entregable comercial fue generado para 32,344 clientes segmentados en seis bandas de prioridad sobre el corte de agosto de 2025. La visualización analítica fue implementada mediante un *dashboard* interactivo en Tableau que permite al equipo comercial explorar los resultados del *scoring* por segmento, banda de prioridad y perfil de portafolio.

La principal contribución del proyecto es demostrar que, incluso bajo condiciones de desbalance extremo de clases (tasa base del 0,121% para 6 meses y del 0,049% para 3 meses sobre el *test* de corte único), un modelo supervisado correctamente diseñado y evaluado mediante métricas de priorización puede producir un entregable comercial de valor operativo, reduciendo el ratio de contacto por conversión de 829:1 (selección aleatoria) a 202:1 (Top 5%) o 139:1 (Top 3%). Las limitaciones del estudio, particularmente el reducido número de eventos positivos y la cobertura parcial de variables explicativas, se detallan en el apartado 6.2.

El enfoque metodológico desarrollado — diseño temporal de cortes con periodo de gap, *pipeline* de preprocesamiento reproducible y evaluación mediante métricas de priorización — es transferible a otros productos del portafolio asegurador de Nacional Seguros.

6.2 Conclusiones personales

Desarrollar este trabajo supuso un reto importante en la unión entre las necesidades del negocio asegurador y las técnicas de analítica avanzada. Trabajar con un *dataset* corporativo de esta complejidad (múltiples líneas de negocio, cortes temporales mensuales, criterios de vigencia) fue una oportunidad para explorar estructuras de datos con las que no había tenido oportunidad de trabajar antes. Aprender a relacionar tablas operativas, a definir la ventana temporal de un *target* y a garantizar la coherencia de cada corte, fueron aprendizajes muy importantes.

Otro de los aprendizajes más valiosos fue reconocer y arreglar un problema de fuga de información temporal (*data leakage*). En las primeras iteraciones los resultados sobre el test se veían extraordinariamente bien, lo que generó una confianza inicial engañosa. Al analizar el porqué de ese comportamiento, se encontró que la partición temporal entre entrenamiento y test no incluía un gap (periodo de separación) equivalente al horizonte del *target*. Dado que la variable objetivo codifica si un cliente contrata Casa Protegida en los N meses posteriores al corte, no tener gap permitía que las ventanas temporales del *target* de los últimos registros de entrenamiento se solapasen con el periodo del test, comprometiendo la independencia de la evaluación. Una vez incorporado el gap correspondiente (6 meses para el horizonte de 6 meses, 3 meses para el de 3 meses), las métricas bajaron a niveles coherentes con la dificultad real del problema. Este proceso refuerza la importancia de diseñar las particiones temporales con rigor y de cuestionar sistemáticamente resultados que parecen demasiado favorables.

A nivel profesional, esta labor ha reafirmado la certeza de que el valor de la analítica avanzada en el mundo asegurador no está solo en la sofisticación del modelo, sino en saber convertir valores estadísticos en herramientas accionables para equipos comerciales. Esto representado

en la generación del entregable con las bandas de prioridad y perfiles de cliente, y la metodología desarrollada aplicable a otros productos del portafolio de Nacional Seguros y a otras problemáticas del negocio que compartan una estructura similar.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

La metodología desarrollada en el presente trabajo constituye una base para extensiones analíticas y operativas que no han podido abordarse dentro del alcance de este TFM. A continuación, se describen las principales líneas de continuación propuestas.

En primer lugar, el *pipeline* actual requiere ejecución manual por parte del analista para cada nuevo corte temporal. Se propone automatizar el proceso de reentrenamiento periódico mediante la programación del *pipeline* en un entorno orquestado, incluyendo la generación automática de cortes SQL, el refresco del *dataset*, el reentrenamiento del modelo y la generación del entregable actualizado, con alertas sobre posibles degradaciones del rendimiento predictivo (model drift).

En segundo lugar, la metodología validada para Casa Protegida es transferible, con ajustes en el diseño de variables y el *target*, a otros productos del portafolio de Nacional Seguros. La extensión a líneas como Vida o Salud permitiría construir un sistema integrado de *scoring* multidimensional, donde cada cliente recibe una estimación de propensión para cada producto disponible y la priorización comercial se realiza considerando el portafolio completo.

En tercer lugar, se propone la incorporación de técnicas de segmentación no supervisada, concretamente mediante *K-Means*, para identificar perfiles de clientes con comportamientos homogéneos y enriquecer el entregable comercial con una dimensión cualitativa complementaria al score individual de propensión.

Por último, se propone ampliar el conjunto de variables explicativas incorporando información de canal de captación, interacciones con la red comercial y datos de reclamaciones históricas. La integración de estas fuentes, condicionada al acceso a sistemas corporativos adicionales, podría incrementar la capacidad predictiva del modelo.

Capítulo 8. REFERENCIAS

Ballings, M., & Van den Poel, D. (2012). Customer event history for churn prediction: How long is long enough? *Expert Systems with Applications*, *39*(18), 13517-13522. https://doi.org/10.1016/j.eswa.2012.07.006

Chapman, p, Clinton, J., & Kerber, R. (2000). *CRISP-DM 1.0: Step-by-Step Data Mining Guide* (SPSS, Ed.). SPSS Inc. https://www.kde.cs.uni-kassel.de/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

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. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning - ICML '06*, 233-240. https://doi.org/10.1145/1143844.1143874

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer International Publishing. https://doi.org/10.1007/978-3-319-98074-4

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

He, H., & Garcia. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. https://doi.org/10.1109/TKDE.2008.239

Kaishev, V. K., Nielsen, J. P., & Thuring, F. (2013). Optimal customer selection for cross-selling of financial services products. *Expert Systems with Applications*, *40*(5), 1748-1757. https://doi.org/10.1016/j.eswa.2012.09.026

Kamakura, W. A., Ramaswami, S. N., & Srivastava, R. K. (1991). Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *International Journal of Research in Marketing*, *8*(4), 329-349. https://doi.org/10.1016/0167-8116(91)90030-B

Kitchens, B., Dobolyi, D., Li, J., & Abbasi, A. (2018). Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data. *Journal of Management Information Systems*, *35*(2), 540-574. https://doi.org/10.1080/07421222.2018.1451957

Li, S., Sun, B., & Wilcox, R. T. (2005). Cross-Selling Sequentially Ordered Products: An Application to Consumer Banking Services. *Journal of Marketing Research*, *42*(2), 233-239. https://doi.org/10.1509/jmkr.42.2.233.62288

Ling, & Li. (1998). Data mining for direct marketing: Problems and solutions. *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98)*, 73-79. https://cdn.aaai.org/KDD/1998/KDD98-011.pdf

Naciones Unidas. (2015). *Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible*. https://unctad.org/system/files/official-document/ares70d1_es.pdf

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592-2602. https://doi.org/10.1016/j.eswa.2008.02.021

Provost, F., & Fawcett, T. (2001). Robust Classification for Imprecise Environments. *Machine Learning*, *42*(3), 203-231. https://doi.org/10.1023/A:1007601015854

Rosset, S., Neumann, E., Eick, U., Vatnik, N., & Idan, I. (2001). Evaluation of prediction models for marketing campaigns. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 456-461. https://doi.org/10.1145/502512.502581

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. https://doi.org/10.1371/journal.pone.0118432

Wirth, & Hipp. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29-39. https://www.semanticscholar.org/paper/48b9293cfd4297f855867ca278f7069abc6a9c24

Capítulo 9. ANEXOS

9.1 Anexo A. Diccionario de datos del dataset gold

El *dataset* gold utilizado en el presente trabajo comprende 1,268,176 registros y 46 columnas, construido a partir de las tablas transaccionales consolidadas de Nacional Seguros. A continuación, se presenta la descripción completa de las variables.

9.1.1 A.1. Variables del dataset

Variable	Tipo	Descripción
FechaCorte	datetime	Fecha del corte temporal de análisis
IdContratante	int64	Identificador único del contratante (PK)
contratante_edad_anios_al_corte	float64	Edad en años al momento del corte
ContratanteGenero	object	Género (Hombre/Mujer/Desconocido)
ContratanteEstadoCivil	object	Estado civil del contratante
flag_tiene_salud_12m	int64	Tiene póliza de Salud activa en 12 meses (0/1)
flag_tiene_vida_12m	int64	Tiene póliza de Vida activa en 12 meses (0/1)
flag_tiene_auto_12m	int64	Tiene póliza de Auto activa en 12 meses (0/1)
salud_monto_ajustado_usd_12m	float64	Suma de primas Salud (USD) en 12 meses
salud_n_polizas_con_prima_12m	int64	Cantidad de pólizas Salud con prima activa
salud_n_polizas_renovadas_12m	int64	Pólizas Salud renovadas en 12 meses
salud_pct_polizas_renovadas_12m	float64	Porcentaje de renovación en Salud
salud_n_asegurados_distintos_12m	int64	Asegurados distintos en pólizas Salud
salud_recency_prima_dias	int64	Días desde última prima de Salud
vida_monto_ajustado_usd_12m	float64	Suma de primas Vida (USD) en 12 meses
vida_n_polizas_con_pago_12m	int64	Pólizas Vida con pago en 12 meses
vida_n_cuotas_pagadas_12m	int64	Cuotas pagadas en Vida en 12 meses
vida_recency_pago_dias	int64	Días desde último pago de Vida
auto_monto_ajustado_usd_12m	float64	Suma de primas Auto (USD) en 12 meses
auto_n_polizas_emitidas_12m	int64	Pólizas Auto emitidas en 12 meses
auto_n_polizas_renovadas_12m	int64	Pólizas Auto renovadas en 12 meses
auto_pct_polizas_renovadas_12m	float64	Porcentaje de renovación en Auto
auto_recency_emision_dias	int64	Días desde última emisión de Auto
antiguedad_any_meses_al_corte	int64	Meses desde primera compra de cualquier producto
cp_compro_6m	int8	*TARGET*: Compró Casa Protegida en 6 meses (0/1)
cp_compro_3m	int8	*TARGET*: Compró Casa Protegida en 3 meses (0/1)

Tabla 16: Descripción de las variables principales del dataset gold. Fuente: Elaboración propia.

Nota: Se presentan las 26 variables principales. El *dataset* completo incluye 46 columnas con variables auxiliares adicionales (fechas de primera compra, montos y conteos complementarios por línea de negocio).

9.2 Anexo B. Variables seleccionadas para el modelo

El *pipeline* de preprocesamiento recibe 35 variables base (33 numéricas y 2 categóricas) y genera 42 variables transformadas tras la aplicación de OneHotEncoder sobre las variables

categóricas (ContratanteGenero con 3 categorías y ContratanteEstadoCivil con 6 categorías). La tabla siguiente resume la importancia de las 20 variables más relevantes del modelo XGBoost optimizado para el horizonte de 6 meses.

#	Variable	Importancia	% Acumulado
1	auto_monto_ajustado_usd_12m	0.0554	5.54%
2	salud_n_asegurados_distintos_12m	0.0550	11.03%
3	flag_tiene_auto_12m	0.0528	16.31%
4	salud_flag_tuvo_anulaciones_12m	0.0502	21.33%
5	vida_n_cuotas_pagadas_12m	0.0430	25.63%
6	salud_antiguedad_meses	0.0342	29.04%
7	salud_monto_ajustado_usd_12m	0.0330	32.35%
8	antiguedad_any_meses_al_corte	0.0328	35.63%
9	vida_recency_pago_dias	0.0315	38.78%
10	vida_monto_ajustado_usd_12m	0.0314	41.92%
11	salud_max_asegurados_por_poliza_12m	0.0310	45.02%
12	ContratanteGenero_Desconocido	0.0299	47.92%
13	auto_n_polizas_emitidas_12m	0.0294	50.87%
14	salud_pct_polizas_renovadas_12m	0.0278	53.64%
15	salud_avg_asegurados_por_poliza_12m	0.0267	56.31%
16	ContratanteEstadoCivil_Soltero(a)	0.0264	58.95%
17	contratante_edad_anios_al_corte	0.0264	61.59%
18	vida_n_polizas_con_pago_12m	0.0256	64.15%
19	flag_tiene_salud_12m	0.0252	66.66%
20	vida_antiguedad_meses	0.0252	69.18%

Tabla 17: Importancia de las 20 variables principales del modelo XGBoost optimizado (horizonte de 6 meses). Fuente: Elaboración propia.

Las 20 variables principales acumulan el 69.18% de la importancia total. El monto de prima en Auto y el número de asegurados distintos en Salud son los predictores dominantes, seguidos por la tenencia de póliza de Auto y el indicador de anulaciones en Salud. Las variables demográficas (género y estado civil) contribuyen de forma secundaria pero significativa.

9.3 Anexo C. Configuración de GridSearchCV

La optimización de hiperparámetros se realizó mediante GridSearchCV con 128 combinaciones evaluadas mediante validación cruzada cronológica de 3 folds (PredefinedSplit). La métrica de selección fue average_precision (PR-AUC).

Hiperparámetro	Valores evaluados
n_estimators	300, 500
learning_rate	0.03; 0.05
max_depth	3, 5
min_child_weight	1, 5
subsample	0.8; 1.0
colsample_bytree	0.8; 1.0
reg_lambda (L2)	1.0; 5.0

La configuración del modelo de 6 meses adoptó mayor profundidad de árboles (max_depth = 5) compensada con submuestreo de filas y columnas, mientras que el modelo de 3 meses mantuvo árboles superficiales (max_depth = 3) sin submuestreo, consistente con un menor volumen de ejemplos positivos disponibles para el entrenamiento.

9.4 Anexo D. Métricas completas del modelo

Métrica	Base (test)	Base (corte único)	Tuned (test)	Tuned (corte único)
n_registros	96,997	32,344	96,997	32,344
n_positivos	118	39	118	39
tasa_base	0.001217	0.001206	0.001217	0.001206
PR-AUC	0.002380	0.003282	0.003327	0.004408
ROC-AUC	0.661941	0.697239	0.726734	0.770062

Tabla 18:Métricas de evaluación del modelo base y optimizado — horizonte de 6 meses. Fuente: Elaboración propia.

Métrica	Base (test)	Base (corte único)	Tuned (test)	Tuned (corte único)
n_registros	97,932	32,694	97,932	32,694
n_positivos	52	16	52	16
tasa_base	0.000531	0.000489	0.000531	0.000489
PR-AUC	0.001006	0.001247	0.001857	0.003891
ROC-AUC	0.636385	0.633751	0.743091	0.75292

Tabla 19: Métricas de evaluación del modelo base y optimizado — horizonte de 3 meses. Fuente: Elaboración propia.

Grupo Top-K	Clientes	Positivos	Conversión	Recall	*Lift*
Top 1%	323	1	0.31%	2.60%	2.57×
Top 3%	970	7	0.72%	17.90%	5.98×
Top 5%	1,617	8	0.49%	20.50%	4.10×
Top 10%	3,234	16	0.49%	41.00%	4.10×
Top 20%	6,468	24	0.37%	61.50%	3.08×

Tabla 20: Análisis Top-K del modelo optimizado de 6 meses sobre el test de corte único (agosto 2025, 32,344 clientes, 39 positivos observados, tasa base 0.12%). Fuente: Elaboración propia.

9.5 Anexo E. Estructura del entregable comercial

El entregable de marketing se materializó en un archivo Excel multi-hoja con la siguiente estructura:

Hoja	Registros	Columnas	Descripción
UniversoScoreado	32,344	20	*Dataset* completo con scores y bandas

ResumenEjecutivo	10	2	Métricas agregadas y KPIs del *scoring*
ResumenBandas	6	5	Análisis por banda de priorización
ResumenTopK	4	8	Análisis Top-K acumulado
Top_0_5	161	20	Clientes en máxima prioridad (Top 0.5%)
Top_1	323	20	Clientes en Top 1% acumulado
Top_5	1,617	20	Clientes en Top 5% acumulado

Tabla 21: Estructura de hojas del entregable comercial (entregable_marketing_scoring.xlsx). Fuente: Elaboración propia.

Cada hoja de detalle (Top_0_5, Top_1, Top_5) incluye: identificador del contratante, score de propensión, ranking, banda de priorización, segmento de prioridad, recomendación de acción comercial, perfil de portafolio, variables demográficas (género, estado civil, edad, antigüedad), flags de tenencia de producto (Salud, Vida, Auto) y montos de prima por línea de negocio.