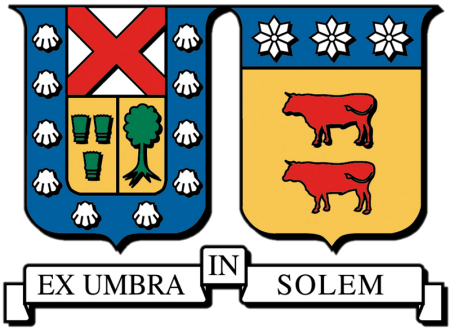


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS

MODELADO PREDICTIVO DE FUGA DE CLIENTES EN EL SECTOR ASEGURADOR CHILENO MEDIANTE MODELOS ECONOMÉTRICOS Y TÉCNICAS DE MACHINE LEARNING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

AUTOR:

NICOLÁS SANTIAGO ABURTO SÁEZ

PROFESOR GUÍA:

ROLANDO RUBILAR TORREALBA

PROFESOR CO-REFERENTE:

BERNARDO PINCHEIRA SARMIENTO

VALPARAÍSO DE CHILE, DICIEMBRE 2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Modelado predictivo de fuga de clientes en el sector asegurador chileno mediante modelos econométricos y técnicas de Machine Learning

Nombre del candidato(a): Nicolás Santiago Aburto Sáez

Carrera / Grado: Ingeniería civil industrial / Pregrado

Campus: Casa Central, Valparaíso. **Departamento:** Industrias

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, _____Rolando Rubilar Torrealba_____, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 9 de diciembre 2025 **Firma:** _____

Estudiante o Candidato(a):

Fecha: 9 de diciembre 2025 **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Tabla de Contenidos

Agradecimientos

Antes de comenzar este escrito, quisiera dar las gracias a todas las personas que fueron importantes en este proceso. En primer lugar, a mi madre y a mi padre, por todo el esfuerzo entregado en mi formación a lo largo de estos años. A mi padre, por ser el pilar de motivación, inteligencia y sustento que me permitió avanzar en esta carrera. A mi madre, por su presencia cálida, su leal compañía y por su esfuerzo incansable en mi crianza, siempre con un corazón amable.

Agradezco también a mi profesor guía, Rolando, por su orientación, paciencia y constante apoyo durante el desarrollo de este trabajo.

Asimismo, a los amigos que hice en el camino —Cristóbal, Isabella, Cassandra, Constanza y Alonso— quienes compartieron conmigo mucho más que jornadas de estudio y se transformaron en un apoyo fundamental en este recorrido, tanto dentro de la universidad como fuera de ella.

Finalmente, a mis más leales amistades de la vida que han estado desde el principio —Catalina, Dominique y Constanza— por haberme recibido en esta ciudad desconocida el 2017 y por convertirse en un apoyo incondicional que ha trascendido los años. Tienen mi cariño y gratitud.

Sin más que agregar, infinitas gracias.

Resumen

El presente trabajo aborda la predicción de fuga de clientes en el sector asegurador chileno mediante la aplicación de modelos econométricos (logit y probit) y técnicas de aprendizaje automático (redes neuronales multicapa). Utilizando una base de datos anonimizada de una aseguradora local, se buscó identificar los factores determinantes de la cancelación de pólizas y comparar el desempeño de ambos enfoques en términos de capacidad predictiva e interpretabilidad.

Los resultados indican que el modelo logit, aunque más limitado en desempeño predictivo, aporta un valor significativo al permitir identificar y cuantificar las variables que influyen en la fuga, como la antigüedad del asegurado, la existencia de documentos impagos y el monto de la prima básica. En cambio, las redes neuronales presentan mejores métricas de evaluación que los modelos econométricos tradicionales en indicadores como *recall*, F1 y AUC, consolidándose como una herramienta eficaz para anticipar clientes en riesgo y habilitar sistemas de alerta temprana.

Se concluye que ambos enfoques no son excluyentes, sino complementarios: el modelo logit aporta capacidad explicativa para la toma de decisiones estratégicas, mientras que la red neuronal mejora la detección de clientes con mayor probabilidad de fuga. Esta combinación ofrece a la aseguradora un marco integral para fortalecer los programas de retención y optimizar los recursos comerciales.

Desde una perspectiva empresarial, el estudio entrega un modelo replicable y de utilidad práctica, con potencial de extenderse a otros productos financieros, a la detección de fraude y a la incorporación de variables macroeconómicas. En conjunto, este trabajo refuerza la importancia de la analítica avanzada como herramienta estratégica para la sostenibilidad y competitividad de la industria aseguradora en Chile.

Abstract

This study examines customer churn prediction in the Chilean insurance sector through the application of econometric models (logit and probit) and machine learning techniques (multilayer neural networks). Using an anonymized database from a local insurer, the study aims to identify the key determinants of policy cancellations and to compare the performance of both approaches in terms of predictive power and interpretability.

The results indicate that the logit model, although more limited in predictive performance, provides significant value by identifying and quantifying the variables that influence churn, such as policyholder tenure, outstanding payments, and premium amount. In contrast, neural networks show better evaluation metrics than econometric models in indicators such as recall, F1, and AUC, establishing themselves as an effective tool for anticipating at-risk customers and enabling early-warning systems.

It is concluded that both approaches are not mutually exclusive but rather complementary: while the logit model offers explanatory insights for strategic decision-making, the neural network improves the identification of customers with a higher probability of churn. This combination provides the insurer with a comprehensive framework to strengthen retention programs and optimize commercial resources.

From a business standpoint, the study delivers a replicable and practical modeling approach with potential extensions to other financial products, fraud detection, and the inclusion of macroeconomic variables. Overall, this work underscores the relevance of advanced analytics as a strategic tool for enhancing the sustainability and competitiveness of Chile's insurance industry.

1. Introducción

En los últimos años, la gestión de clientes se ha transformado en un eje estratégico en industrias intensivas en servicios, entre ellas el sector asegurador, dado su impacto en ingresos, fidelización y costos de adquisición (Verbeke et al., 2011; Henao Madrigal et al., 2020). La creciente competencia, la presión regulatoria y la irrupción de actores digitales han hecho que la retención de asegurados sea un factor crítico para la sostenibilidad de las compañías (European Insurance and Occupational Pensions Authority (EIOPA), 2024; Organisation for Economic Co-operation and Development (OECD), 2023). En este escenario, anticipar la fuga de clientes permite resguardar ingresos futuros, optimizar los recursos destinados a programas de fidelización y reducir los elevados costos asociados a la adquisición de nuevos clientes.

A nivel internacional, la literatura muestra un tránsito progresivo desde metodologías estadísticas tradicionales hacia enfoques de *machine learning*, impulsado por la necesidad de mejorar la capacidad predictiva frente a patrones de comportamiento cada vez más complejos, lo que a su vez reaviva el debate entre desempeño e interpretabilidad (Breiman, 2001; Rudin et al., 2022). Sin embargo, este avance ha puesto de relieve un dilema central: los modelos más precisos suelen ser también los menos interpretables, lo que dificulta su adopción en sectores regulados donde la transparencia es indispensable. Esta tensión entre la capacidad explicativa y el desempeño predictivo constituye el marco conceptual de este estudio.

Bajo este contexto, el presente trabajo se plantea como un esfuerzo por evaluar y comparar distintos enfoques predictivos aplicados a la realidad del mercado asegurador chileno, buscando aportar evidencia empírica que permita comprender mejor los factores que inciden en la fuga de clientes y fortalecer las estrategias de retención basadas en datos. Así, se pretende no solo avanzar

en la comprensión teórica del fenómeno, sino también entregar herramientas prácticas que contribuyan a la toma de decisiones informadas en un entorno de alta competencia y cambio tecnológico constante. Por razones de confidencialidad, los datos utilizados provienen de una aseguradora nacional y se entregaron en un formato completamente anonimizado; no se incluye el nombre de la entidad ni información sensible de los asegurados, trabajando únicamente con variables y estructuras de datos previamente depuradas y sin posibilidad de identificación individual.

Por otro lado, en cuanto a las motivaciones de este trabajo, se tiene que en Chile, pese a la relevancia del tema, la evidencia empírica sobre predicción de fuga en seguros es aún escasa. La mayoría de los estudios se concentra en sectores como telecomunicaciones o banca, donde históricamente se ha desarrollado la mayor parte de la literatura en abandono de clientes (Lemmens and Croux, 2006). Ello refuerza la necesidad de generar investigaciones específicas para el mercado asegurador, considerando sus particularidades contractuales y regulatorias. Esta brecha abre la oportunidad de producir evidencia local que aporte no solo al desarrollo académico, sino también a la gestión empresarial concreta de las aseguradoras (Lorca Figueroa, 2021).

La presente memoria surge de esa necesidad y busca contrastar dos enfoques: los modelos econométricos clásicos (logit y probit), reconocidos por su valor explicativo, y las redes neuronales multicapa, destacadas por su capacidad predictiva. El objetivo es evaluar cómo cada uno de estos modelos contribuye a la comprensión y anticipación de la fuga de clientes, y de qué manera su uso conjunto puede resultar en una combinación que fortalezca la retención en una aseguradora chilena.

2. Problema de investigación

En un entorno macroeconómico exigente, el mercado asegurador chileno ha mostrado resiliencia, con un crecimiento moderado en el segmento de vida y fundamentos que respaldan una perspectiva estable para el próximo período. En particular, AM Best (2025) mantiene una visión positiva para la industria local, destacando el dinamismo de las primas y los avances regulatorios, mientras que la Comisión para el Mercado Financiero (CMF) (2024) reporta que, a septiembre de 2024, las ventas del mercado crecieron en términos reales (vida: +2,2 %, mercado total: +1,1 % en enero–septiembre). En este contexto, la retención de clientes adquiere un valor estratégico: distintos análisis gerenciales sostienen que captar un nuevo cliente puede costar entre cinco y veinticinco veces más que retener a uno existente, y que aumentos moderados en la retención pueden traducirse en mejoras sustantivas de rentabilidad[1] (Gallo, 2014; Bain & Company, 2014).

Pese a la relevancia que la fidelización tiene para la estabilidad financiera y la planificación comercial, la evidencia aplicada al sector asegurador chileno en materia de predicción de cancelaciones sigue siendo limitada en comparación con otras industrias, como las telecomunicaciones o la banca. En este contexto, la aseguradora busca anticipar con suficiente antelación la cancelación de pólizas, con el fin de orientar sus esfuerzos de retención hacia los clientes con mayor riesgo y así optimizar el uso de los recursos comerciales.

El desafío central radica en predecir, con un grado razonable de exactitud, qué clientes presentan mayor probabilidad de anular o no renovar su póliza, de modo que la empresa pueda actuar preventivamente. Estudios locales, como el de Lorca Figueroa (2021), señalan que los seguros masivos vinculados a la banca presentan tasas de abandono inicial elevadas, aunque condicionadas al

[1] Las magnitudes dependen del sector, el modelo de negocio y la metodología empleada. Se citan estimaciones ejecutivas ampliamente difundidas en gestión comercial.

canal y segmento analizado. Esto refuerza la necesidad de comprender, a partir de los propios datos de la aseguradora, qué factores explican la decisión de fuga y cómo dichas variables interactúan entre sí.

En este sentido, surgen varias inquietudes que orientan el desarrollo de esta investigación. En primer lugar, se busca identificar las variables que influyen significativamente en la probabilidad de cancelación de una póliza, aportando información relevante para la gestión comercial y la toma de decisiones estratégicas. Asimismo, se examina el alcance de los modelos econométricos tradicionales, como logit y probit, en su capacidad para capturar relaciones sustantivas entre las variables y generar interpretaciones útiles para la gestión del negocio. En paralelo, se evalúa el potencial de las técnicas de aprendizaje automático, en particular de las redes neuronales, para mejorar el desempeño predictivo sin renunciar completamente a la interpretación de los resultados. Finalmente, se analiza cómo la combinación de ambos enfoques puede contribuir al diseño de estrategias de retención más efectivas y sostenibles en el tiempo.

El estudio se desarrolla a partir de una base de datos anonimizada a nivel de póliza, que incluye un identificador interno y una variable binaria de anulación (FUGA). Este conjunto de información permite analizar patrones relevantes y construir modelos que apoyen la toma de decisiones estratégicas. Con ello, no solo se busca desarrollar un modelo predictivo con utilidad inmediata, sino también sentar las bases de un sistema de retención replicable que la aseguradora pueda aplicar en futuras líneas de negocio o cohortes de clientes. En definitiva, el propósito último es fortalecer la sostenibilidad operativa y financiera de la empresa, mejorar la relación con los asegurados y aportar evidencia práctica que contribuya a la gestión de la fuga en el mercado asegurador chileno.

3. Objetivos

Objetivo General

Implementar un modelo predictivo de fuga de clientes en una empresa aseguradora chilena, utilizando inicialmente modelos econométricos binarios (logit y probit) y posteriormente comparando su desempeño con técnicas de *machine learning*, en particular redes neuronales, con el propósito de identificar tempranamente a los asegurados con mayor probabilidad de anulación o no renovación, apoyando así la toma de decisiones en estrategias de retención.

Objetivos Específicos

- Desarrollar una caracterización integral de la base de datos de clientes y pólizas, identificando y evaluando las variables relevantes asociadas al comportamiento de fuga.
- Estimar modelos logit y probit para cuantificar el efecto de las principales variables explicativas sobre la probabilidad de fuga, interpretando los efectos marginales y su significancia económica.
- Implementar y entrenar una red neuronal multicapa como alternativa no lineal para la predicción de fuga, optimizando su arquitectura e hiperparámetros a fin de mejorar la capacidad predictiva y contrastar su desempeño con los modelos econométricos.
- Validar el modelo predictivo mediante métricas de desempeño y balance de clases (AUC, precisión, *recall*, entre otras), comparando los resultados de los enfoques econométricos con los obtenidos a través de redes neuronales, con el fin de determinar el modelo con mayor capacidad explicativa y valor práctico para la gestión de retención.

4. Marco teórico

La fuga de clientes, también denominada *customer churn*, se refiere a la decisión voluntaria de un asegurado de no renovar o cancelar anticipadamente su póliza. En la industria aseguradora este fenómeno reviste especial relevancia, pues impacta directamente en la estabilidad financiera y en la proyección de ingresos de las compañías. Diversos estudios han demostrado que retener a un cliente resulta más rentable que captar uno nuevo, tanto en términos de costos como de sostenibilidad de largo plazo (Reichheld and Sasser, 1990; Gallo, 2014).

En Chile no existen estadísticas oficiales de "tasas de fuga" publicadas por la Comisión para el Mercado Financiero (CMF), ya que los reportes sectoriales se concentran en primas, siniestralidad y resultados financieros (Comisión para el Mercado Financiero (CMF), 2024). Sin embargo, estudios internacionales muestran que la tasa de abandono en servicios financieros suele rondar el 19 %, y que en promedio las industrias pierden entre un 10 % y un 25 % de su base de clientes anualmente (CustomerGauge, 2024). Este dinamismo, junto con la irrupción de nuevos competidores digitales, configura un entorno altamente competitivo en el que anticipar la fuga de clientes se convierte en un factor crítico de éxito (AM Best, 2025).

Además, como ya se mencionó anteriormente, la mayor parte de la literatura académica en predicción de fuga (o "churn") se ha focalizado en sectores como telecomunicaciones o banca (industrias con alta rotación de clientes y abundantes datos públicos) en contraste con el sector asegurador, donde la evidencia es mucho más limitada (Lemmens and Croux, 2006). Por su parte, el ámbito asegurador presenta particularidades propias, como la estructura contractual de largo plazo, la dependencia de variables actuariales y las exigencias regulatorias, que justifican la necesidad de investigaciones específicas. En este sentido, abordar el estudio de la fuga en el mercado asegurador

chileno no solo permite adaptar metodologías probadas a un nuevo contexto, sino también generar evidencia empírica local que contribuya a la comprensión del comportamiento de los asegurados y a la formulación de estrategias efectivas de retención.

En función de esta brecha, resulta pertinente revisar las principales metodologías empleadas para modelar la fuga de clientes en la literatura internacional. Entre los enfoques más tradicionales destacan los modelos econométricos de tipo binario, ampliamente utilizados como punto de partida en estudios de retención por su capacidad explicativa y su facilidad para interpretar los factores que influyen en la probabilidad de cancelación.

4.1. Modelos clásicos y limitaciones

Entre las metodologías econométricas más empleadas para modelar variables binarias, como la decisión de fuga o permanencia, destacan los modelos *logit* y *probit*. Ambos pertenecen a la familia de los Modelos Lineales Generalizados (GLM, por sus siglas en inglés), un marco estadístico que extiende la regresión lineal clásica al permitir que la variable dependiente siga distribuciones distintas a la normal y que la relación entre el valor esperado y las covariables sea no lineal (Wooldridge, 2009).

En la industria aseguradora, estos modelos se han consolidado como herramientas de referencia para analizar el comportamiento de cancelación de pólizas o *churn*. Un ejemplo relevante es el trabajo de Dong et al. (2022), quienes emplean una extensión multinomial del modelo *logit* para estudiar la probabilidad de transición entre distintos estados de relación cliente-compañía en seguros generales, mostrando que la historia previa de interacciones tiene un impacto significativo en la probabilidad de retención. En el mismo sentido, Azzone et al. (2022) aplican regresiones logísticas

para predecir la deserción en seguros de vida y concluyen que, aunque las técnicas de *machine learning* logran una mejora marginal en precisión, la regresión logística sigue ofreciendo ventajas interpretativas para la toma de decisiones comerciales. Por su parte, Manteigas and António (2024) confirman esta idea en el segmento de seguros de vida asociados a hipotecas, al comparar modelos de regresión con algoritmos de árboles y mostrar que los coeficientes de la regresión permiten identificar de manera clara las variables críticas vinculadas a la fuga.

Estudios previos en Latinoamérica refuerzan esta vigencia. En particular, Henao Madrigal et al. (2020) modelan la fuga de clientes multiproducto en una aseguradora regional mediante modelos de duración (basados en hazard rates, esto es tasas de riesgo o modelos de duración), pero incorporan regresiones logísticas como contraste, destacando su utilidad para estimar la probabilidad de cancelación temprana de productos. De forma complementaria, análisis recientes de la industria como el informe de InsuredMine (2024) muestran que las aseguradoras continúan utilizando modelos logit y probit como línea base para sus sistemas de predicción, especialmente en etapas iniciales de implementación de analítica de datos.

A pesar de sus virtudes, estos modelos presentan limitaciones cuando se enfrentan a bases de datos extensas y altamente no lineales, como suele ocurrir en contextos de aseguramiento. Su estructura funcional lineal en los parámetros impide capturar interacciones complejas entre variables, como el efecto conjunto de la siniestralidad y la antigüedad del cliente sobre la probabilidad de anulación. Además, cuando el fenómeno de interés es poco frecuente, por ejemplo, una tasa de fuga inferior al 10 %, los modelos tienden a subestimar la clase minoritaria si no se aplican correcciones por desbalance (Burez and Van den Poel, 2009). Esta limitación se acentúa al considerar la evolución temporal del comportamiento de los asegurados, fenómeno conocido como *concept drift*, que puede hacer que un modelo ajustado con datos históricos pierda capacidad predictiva en

contextos cambiantes.

En respuesta a estos desafíos, la literatura ha propuesto diversas extensiones del marco GLM, entre ellas los modelos con enlaces sesgados o asimétricos, que buscan mejorar el ajuste en presencia de respuestas raras o distribuciones fuertemente desbalanceadas. Yin et al. (2020) introducen un modelo de enlace sesgado para respuestas binarias en seguros de vida, mostrando que estas funciones permiten reducir el sesgo y mejorar la calibración en predicciones de eventos infrecuentes, como la cancelación anticipada o la no renovación de pólizas.

Pese a lo anterior, los modelos *logit* y *probit* siguen constituyendo una buena base metodológica en estudios de retención y fuga en seguros. Su principal fortaleza radica en la claridad interpretativa y la facilidad para comunicar resultados a nivel ejecutivo, mientras que sus limitaciones motivan la incorporación de enfoques más flexibles, como los modelos de *machine learning*, capaces de capturar relaciones más complejas entre variables.

4.2. Avances en machine learning

Durante la última década, el uso de algoritmos de *machine learning* (ML) para predecir la fuga de clientes se ha expandido con fuerza en sectores intensivos en datos, como seguros, banca y telecomunicaciones. Estos métodos destacan por su capacidad para capturar relaciones no lineales y estructuras complejas que los modelos clásicos difícilmente representan. En un estudio comparativo a gran escala, Bogaert and Delaere (2023) evaluaron 33 clasificadores en 11 conjuntos de datos de *churn*, demostrando que los modelos de ensamble heterogéneo superan de manera sistemática a los clasificadores individuales, lo que confirma la ventaja de combinar múltiples inductores en contextos de alta dimensionalidad.

En el ámbito asegurador, los métodos basados en árboles potenciados han adquirido especial relevancia por su flexibilidad y poder explicativo. Manteigas and António (2024) evidencian que algoritmos como XGBoost logran capturar con mayor precisión la relación entre variables de comportamiento y la probabilidad de anulación en seguros de vida hipotecarios. De forma complementaria, AbdelAziz et al. (2025) comparan XGBoost, una CNN y un ensamble profundo en distintos sectores, incluido uno de seguros, reportando desempeños sobresalientes en exactitud y F1, lo que refuerza la aplicabilidad transversal de estas técnicas.

Uno de los principales retos en predicción de fuga es el desbalance de clases, dado que la cantidad de clientes que cancelan suele ser mucho menor que la de quienes permanecen. Para abordar este problema, la literatura recomienda estrategias como el sobremuestreo sintético (SMOTE) (Chawla et al., 2002) o su variante adaptativa ADASYN (He et al., 2008), que generan observaciones artificiales de la clase minoritaria para mejorar el aprendizaje. Estas técnicas, combinadas con validación cruzada estratificada y métricas específicas, permiten equilibrar el desempeño entre sensibilidad y precisión (Burez and Van den Poel, 2009). Un caso reciente es el de Liu et al. (2024), quienes integran ADASYN con una arquitectura híbrida de atención, BiLSTM y CNN, obteniendo mejoras significativas frente a modelos tradicionales en datos de seguros, banca y telecomunicaciones.

Los enfoques más recientes también incorporan la evolución temporal de los asegurados mediante modelos longitudinales y multietapa. Valla et al. (2024) proponen un marco basado en árboles y bosques para analizar datos con censura o truncamiento, demostrando que incorporar la trayectoria del cliente mejora la predicción y la toma de decisiones sobre retención. Asimismo, el uso de métodos parcimoniosos como Lasso permite identificar factores determinantes de la anulación manteniendo interpretabilidad (Reck et al., 2023).

De esta manera, los avances en ML han elevado la precisión y robustez de los modelos de fuga al permitir una mayor adaptación a datos desbalanceados y relaciones no lineales. Sin embargo, su éxito depende de prácticas rigurosas de validación y calibración probabilística. En consecuencia, la literatura coincide en que los modelos clásicos, como logit y probit, siguen siendo fundamentales como punto de partida interpretativo, mientras que los métodos de *boosting*, ensambles y arquitecturas profundas constituyen una evolución natural cuando el objetivo es maximizar la capacidad predictiva en escenarios complejos del sector asegurador. Esta transición hacia modelos más sofisticados abre paso a un campo que ha cobrado un protagonismo creciente en los últimos años: las redes neuronales.

4.3. Redes neuronales

Las redes neuronales han ganado protagonismo en la predicción de *churn* debido a su capacidad para aprender relaciones complejas y no lineales entre un gran número de variables. A diferencia de los modelos clásicos, las redes pueden identificar patrones ocultos en los datos y procesar secuencias de información, como la evolución temporal del comportamiento de un cliente. Dentro de este enfoque, se han desarrollado arquitecturas más avanzadas, conocidas como modelos híbridos, que combinan distintos tipos de redes para aprovechar sus fortalezas.

Uno de los ejemplos más utilizados son las redes neuronales convolucionales (CNN, por sus siglas en inglés *Convolutional Neural Networks*), diseñadas originalmente para procesar información con estructura espacial, pero que también se aplican en análisis de series de tiempo y comportamiento de clientes. Estas pueden complementarse con redes de memoria a largo y corto plazo bidireccionales (BiLSTM, *Bidirectional Long Short-Term Memory*), las cuales son un tipo

de red recurrente capaz de aprender dependencias temporales, es decir, cómo las acciones pasadas de un cliente influyen en su comportamiento futuro. En conjunto, estos modelos logran capturar tanto la secuencia temporal como las interacciones entre variables, lo que se traduce en mejores resultados predictivos.

De hecho, Liu et al. (2024) proponen una arquitectura híbrida que combina mecanismos de atención, BiLSTM y CNN, alcanzando mejoras significativas en métricas como precisión, *recall* y F1 en contextos de seguros, banca y telecomunicaciones. De forma similar, AbdelAziz et al. (2025) muestran que los modelos de redes neuronales profundas (conocidos como *deep learning*) y los ensambles de arquitecturas avanzadas pueden superar a los algoritmos de *boosting* tradicionales, manteniendo resultados estables en distintos sectores.

No obstante, la implementación de estas redes implica nuevos desafíos. Entre ellos se encuentran el riesgo de sobreajuste, es decir, cuando el modelo aprende demasiado los datos históricos y pierde capacidad de generalización, la sensibilidad a los hiperparámetros, que son los valores que controlan el entrenamiento del modelo, y los problemas de interpretabilidad, ya que las redes complejas suelen funcionar como “cajas negras”. Por esta razón, en industrias reguladas como la aseguradora, su uso se complementa con validaciones orientadas al negocio, donde se consideran métricas como el costo de los errores, la ganancia esperada o la sensibilidad a diferentes umbrales de predicción. Todo esto permite garantizar un uso responsable, transparente y alineado con los objetivos de retención de clientes.

4.4. Métricas de evaluación

Por último, para poder comparar y validar los enfoques descritos anteriormente, resulta esencial definir métricas de evaluación adecuadas que permitan medir objetivamente el desempeño predictivo de los modelos. En la literatura sobre fuga de clientes, la correcta selección de métricas ha cobrado gran relevancia, especialmente debido al desbalance de clases característico de este tipo de problemas, donde el número de clientes que permanecen supera ampliamente al de los que se fugan.

Los estudios recientes sobre *churn prediction* coinciden en la utilización de métricas como la precisión, el *recall*, la puntuación F1 y el área bajo la curva ROC (AUC-ROC) (Bogaert and Delaere, 2023; AbdelAziz et al., 2025; Liu et al., 2024). Cada una de estas métricas entrega información complementaria. La precisión refleja la proporción de predicciones correctas entre los casos clasificados como fuga, mientras que el *recall* (o sensibilidad) mide la capacidad del modelo para identificar correctamente a los clientes que efectivamente abandonan la compañía. La puntuación F1 corresponde a la media armónica entre precisión y *recall*, y se utiliza cuando se requiere equilibrar ambos indicadores. Por su parte, el AUC-ROC evalúa la capacidad global del modelo para discriminar entre clientes que se fugan y los que permanecen, siendo especialmente útil cuando las clases están desbalanceadas.

En el contexto asegurador, la elección de estas métricas depende del objetivo operativo y de los costos asociados a los errores de clasificación. Por ejemplo, maximizar el *recall* es recomendable cuando el costo de no detectar a un cliente en riesgo (falso negativo) es alto, mientras que priorizar la precisión es más adecuado si los recursos destinados a retención son limitados y se busca evitar intervenciones innecesarias. De esta forma, el uso conjunto de estas métricas

permite obtener una visión más equilibrada y realista del desempeño del modelo, facilitando la comparación entre distintos enfoques y su posterior aplicación práctica en estrategias de retención.

5. Metodología

5.1. Descripción de la base de datos

La base de datos utilizada corresponde a registros administrativos de una empresa aseguradora en Chile.[2] El conjunto contiene **300,984 observaciones** (póliza–cliente, según la definición operacional de la compañía) y se construyó para estudiar la fuga de clientes entendida como *anulación o no renovación* de la póliza. La variable dependiente es binaria: $FUGA{=}1$ si la póliza se anula/no renueva y $FUGA{=}0$ si permanece vigente.

NRPOLI	FUGA	EDAD	SEXO	ANTIG. (mes)	MNPRBA	MNPMP	IMPAGOS	…	INTERMED_A	ZONA_V
11000001	1	38	1	26	0.51	0.1564	1	…	1	0
11000002	1	39	1	135	1.50	1.1449	0	…	0	0
11000003	0	48	0	197	1.00	0.6187	6	…	0	0
11000005	1	24	0	110	1.00	0.6554	1	…	1	1
11000009	0	24	0	197	0.50	0.1554	0	…	0	0
11000012	1	25	0	158	0.5128	0.1680	0	…	0	0
11000013	1	45	0	147	0.6804	0.3110	94	…	0	0
11000014	1	35	0	191	2.00	1.6499	2	…	0	0
11000015	1	39	0	126	0.3568	0.0017	1	…	0	0

Tabla 1: Muestra representativa de los datos anonimizados utilizados en el estudio. Se presenta un extracto de nueve filas provenientes de la base de registros, con el fin de ilustrar la estructura original y el formato en que se encuentran disponibles las variables. Se omiten columnas intermedias para facilitar su visualización.

Las variables explicativas incluyen:

- **Cuantitativas**: `EDAD`, `ANTIGUEDAD_MESES`, `DOC_IMPAGOS`, `MNPRBA` (prima básica), `MNCAAS` (capital asegurado), `Nueva_MNPMP` (prima excedente), y `BENEFICIOS_RUT`.
- **Categóricas (dummies)**: `SEGURO_`$\star$ (tipo de producto, $\sim$43 indicadores), `INTERMEDIARIO_`$\star$ (canal de venta, $\sim$8 indicadores) y `ZONA_`$\star$ (sucursal/zona, $\sim$22 indicadores).

[2]Por acuerdo de confidencialidad, se omite el nombre comercial y cualquier dato que permita identificar a clientes o productos específicos.

Este diseño es consistente con un modelo de clasificación binaria (logit/probit) estimado por máxima verosimilitud, lo que garantiza probabilidades en $[0, 1]$ y permite efectos marginales no constantes. Véase Wooldridge (2009) para la derivación de la log–verosimilitud y propiedades de EMV, y StataCorp LLC (2025) para la formulación y opciones de estimación reportadas por el software.

Este conjunto de variables se obtuvo a partir de un proceso previo de depuración y transformación de la base de datos original, que incluyó el tratamiento de valores faltantes, la construcción de variables derivadas y la codificación de las variables categóricas en formato *dummy*. El detalle completo de las rutinas de limpieza y preparación de datos, implementadas en `Python`, se presenta en el Anexo B.

5.2. Modelos Logit y Probit

En este trabajo se busca modelar la probabilidad de fuga de clientes, entendida como la anulación o no renovación de una póliza de seguros. Para ello se utilizan modelos de regresión para variables dependientes binarias, siendo los más comunes el *modelo logit* y el *modelo probit*.

Estos modelos permiten superar las limitaciones del Modelo de Probabilidad Lineal (MPL), ya que garantizan que las probabilidades predichas se encuentren en el rango $[0, 1]$ y permiten que los efectos marginales de las variables explicativas no sean constantes, sino que dependan de los valores de los regresores (Wooldridge, 2009).

La especificación general de estos modelos se basa en lo siguiente:

$$P(FUGA_i = 1 \mid X_i) = G(X_i\beta), \tag{1}$$

donde $FUGA_i$ es la variable dependiente que toma valor 1 si la póliza es anulada y 0 en caso contrario; X_i es el vector de variables explicativas asociadas a la póliza i; β es el vector de parámetros a estimar; y $G(\cdot)$ corresponde a una función de distribución acumulada (CDF).

- Si $G(\cdot)$ corresponde a la función logística estándar, se obtiene el **modelo logit**:

$$G(z) = \frac{e^z}{1+e^z}. \tag{2}$$

- Si $G(\cdot)$ corresponde a la función de distribución normal estándar, se obtiene el **modelo probit**:

$$G(z) = \Phi(z), \tag{3}$$

  donde $\Phi(\cdot)$ representa la CDF de una normal estándar.

La estimación de ambos modelos se realiza mediante el método de **Máxima Verosimilitud (MV)**. En el caso del modelo logit, la función de log-verosimilitud está dada por (StataCorp LLC, 2025):

$$\ln L = \sum_{j \in S} w_j \ln F(x_j, \beta) \; + \; \sum_{j \notin S} w_j \ln\{1 - F(x_j, \beta)\}, \tag{4}$$

donde $F(x_j, \beta)$ corresponde a la probabilidad estimada de fuga para la observación j, S es el conjunto de pólizas anuladas ($FUGA = 1$), y w_j son los posibles pesos asociados a cada observación.

El interés principal de este tipo de modelos no recae en el valor directo de los coeficientes β, ya que no representan cambios marginales lineales en la probabilidad. Lo relevante son:

- El **signo de los coeficientes**, que indica si una variable aumenta o disminuye la probabilidad

de fuga.

- Los **efectos marginales**, que cuantifican el cambio en la probabilidad de fuga ante una variación en la variable explicativa, manteniendo constantes las demás.

De este modo, el modelo logit/probit permite responder preguntas de negocio directamente relacionadas con la gestión de clientes, como por ejemplo: "¿Cuánto aumenta la probabilidad de fuga si un cliente acumula documentos impagos?" o "¿Cuál es el impacto de la antigüedad de la póliza en la probabilidad de cancelación?".

Efectos marginales

Dado que los coeficientes estimados en modelos logit y probit no representan cambios directos en la probabilidad, la interpretación práctica se realiza a través de los **efectos marginales**. En términos generales, el efecto marginal de la variable x_j sobre la probabilidad de fuga está dado por:

$$\frac{\partial P(FUGA = 1 \mid X)}{\partial x_j} = g(X\beta) \cdot \beta_j, \tag{5}$$

donde $g(\cdot)$ corresponde a la función de densidad asociada a $G(\cdot)$: en el caso del logit es la función logística $g(z) = \frac{e^z}{(1+e^z)^2}$ y en el caso del probit es la densidad normal estándar $\phi(z)$.

En la práctica, estos efectos marginales se suelen evaluar en el promedio de la muestra (Marginal Effect at the Mean, MEM) o promediando el efecto individual de cada observación (Average Marginal Effect, AME). Este último enfoque, calculado mediante el comando `margins` en el software estadístico *Stata* o de forma equivalente en el lenguaje de programación *Python*, utilizando librerías especializadas como `statsmodels`, mediante el método `get_margeff()`,

o el paquete `marginspy`. En este trabajo se reportan los efectos marginales promedio (AME), ya que ofrecen una medida representativa del cambio esperado en la probabilidad de fuga frente a variaciones en cada variable explicativa.

5.2.1. Estadísticos descriptivos de variables principales

Antes de estimar, se reportan estadísticos descriptivos de un subconjunto de variables *relevantes para el fenómeno de fuga*. Esta tabla se actualizará tras la etapa de selección y validación (por ejemplo, usando efectos marginales del logit y medidas de importancia en el clasificador de *machine learning*).

Tabla 2: Estadísticos descriptivos de variables clave

Variable	Media	Desv. Est.	Mín.	Máx.	Notas
FUGA (0/1)	0.564	0.496	0	1	Indicador de anulación/no renovación.
MNCAAS	177.711	255.869	0	5000	Capital asegurado.
MNPRBA	0.880	0.863	0.008	46.6	Prima básica mensual.
EDAD	39.232	11.086	18	87	Edad del asegurado titular.
SEXO (0/1)	0.629	0.483	0	1	Género del asegurado: 0 = masculino, 1 = femenino.
DOC_IMPAGOS	2.157	4.907	0	177	Nº de documentos impagos a la fecha de corte.
ANTIGUEDAD_MESES	36.261	35.100	0	197	Tiempo desde emisión/alta de póliza.

En el contexto de fuga, `FUGA` resume el evento de interés; `MNPRBA` y `MNCAAS` reflejan el tamaño económico del contrato; `EDAD` permite capturar diferencias demográficas; `SEXO` permite controlar posibles diferencias de comportamiento por género; `DOC_IMPAGOS` captura la tensión de pago que puede gatillar la anulación de pólizas y `ANTIGUEDAD_MESES` aproxima la madurez de la relación asegurado–compañía.

5.2.2. Análisis descriptivo e implicancias para el modelado

FUGA (0/1)

La media de $0{,}564$ indica una *prevalencia* de fuga del 56,4 %. Al ser una variable Bernoulli, su desviación estándar teórica es $\sqrt{p(1-p)}$, que con $p{=}0{,}564$ da $\approx\ 0{,}495$, consistente con el valor observado $(0{,}496)$. Esto sugiere que la clase positiva *no es rara* en esta muestra; por lo tanto, la evaluación del clasificador debe considerar una línea base ingenua cercana a 56 % (p. ej., "siempre predecir fuga"). En este contexto, métricas como *balanced accuracy*, F1 y AUC–PR seguirán siendo informativas, pero la discusión de umbrales y calibración será igual de relevante que la métrica agregada.

MNCAAS (capital asegurado)

Promedio $177{,}71$ con DE $255{,}87$ y máximo 5000 revela marcada asimetría y una cola larga. El valor mínimo 0 no corresponde a un error de registro, sino que refleja a los productos de *ahorro e inversión*, en los cuales el asegurado no define un capital específico, sino que busca acumular recursos a través del tiempo. En este sentido, los ceros son parte de la heterogeneidad del portafolio y deben conservarse. De todos modos, al igual que en la prima, una transformación *log(1+x)* puede ser útil para estabilizar la relación con la probabilidad de fuga y mejorar la interpretación marginal en presencia de distribuciones sesgadas.

MNPRBA (prima básica)

Media $0{,}880$ y DE $0{,}863$ con rango $[0{,}008,\ 46{,}6]$ evidencian fuerte asimetría a la derecha (colas largas). Esta heterogeneidad por tamaño del contrato sugiere considerar transformaciones

log (p. ej., $\log(1+\text{MNPRBA})$) para estabilizar la relación con la probabilidad de fuga y reducir la influencia de valores extremos en la estimación por máxima verosimilitud.

EDAD

Presenta una media de 39,23 años (DE 11,09) y un rango amplio (18 a 87). La dispersión sugiere perfiles etarios heterogéneos y, en términos de relación con la fuga, es plausible una *no linealidad* (por ejemplo, riesgos distintos en edades bajas vs. altas). En el logit/probit se puede permitir curvatura con un término cuadrático (EDAD^2) o con funciones flexibles (splines), lo que evaluaremos empíricamente.

SEXO (0 masculino, 1 femenino)

La media de $0{,}629$ indica que el 62,9 % de la muestra corresponde al código 1 (femenino). Dado que existen líneas de producto con segmentación por género, esta variable funcionará principalmente como *control*. En la etapa de modelación conviene verificar posibles *interacciones* con dummies de producto (`SEGURO_*`) o canal (`INTERMEDIARIO_*`) si hubiese evidencia de efectos diferenciales.

DOC_IMPAGOS

La media de $2{,}157$ y DE $4{,}907$ (máximo 177) confirman una distribución fuertemente asimétrica con cola larga y muchos ceros (clientes al día). Dado su rol operativo, es razonable esperar una asociación positiva con la fuga. Para el logit/probit conviene probar especificaciones con transformaciones *log(1+x)* o discretizaciones (p. ej., indicadores ≥ 1, ≥ 3, ≥ 6 impagos) y comparar efectos marginales promedio (AME).

ANTIGUEDAD_MESES

La media de 36,26 meses (DE 35,10) y un máximo de 197 meses muestran una mezcla de pólizas muy nuevas y muy antiguas. Es común encontrar patrones no lineales (p. ej., mayor riesgo al inicio, estabilización después), por lo que también aquí conviene permitir curvatura (término cuadrático o splines) o discretizaciones por tramos (cohortes de antigüedad) en ejercicios de robustez.

Los descriptivos muestran (i) una clase positiva relativamente frecuente, lo que facilita la estimación pero exige discutir umbrales de decisión; (ii) variables económicas con *colas largas* (MNPRBA y MNCAAS), candidatas a transformaciones logarítmicas; (iii) covariables con probable *no linealidad* (EDAD, ANTIGUEDAD_MESES), para las que se recomienda permitir curvatura; y (iv) un predictor operativo clave con distribución altamente sesgada (DOC_IMPAGOS), para el que conviene evaluar transformaciones o puntos de corte. Estas consideraciones se incorporarán en la especificación del logit/probit y en la comparación posterior con el clasificador de *deep learning*, reportando *efectos marginales promedio* (AME) para interpretar la magnitud de los cambios en la probabilidad de fuga.

En línea con estas evidencias descriptivas, la sección siguiente fija la especificación del modelo (transformaciones logarítmicas en montos, término cuadrático en edad con centrado, y el bloque de dummies de control), así como los chequeos de identificabilidad y robustez.

5.3. Especificaciones del modelo

Se estima un logit sobre la misma muestra, con: (i) transformaciones $\log(1+x)$ en `MNCAAS` y `MNPRBA`; (ii) `EDAD` centrada en su media e inclusión de $\texttt{EDAD}^2$ para permitir curvatura; y (iii) el

mismo bloque de dummies de control por producto, canal y zona (`SEGURO_*`, `INTERMEDIARIO_*`, `ZONA_*`). Se limpian automáticamente dummies sin variación o con predicción perfecta.[3].

Las especificaciones a considerar son:

- **Modelo 1:** Incluye únicamente las variables económicas principales: $\log(1 + \texttt{MNCAAS})$ y $\log(1 + \texttt{MNPRBA})$.

$$P(\texttt{FUGA}_i = 1 \mid X_i) = G\Big(\beta_0 + \beta_1 \log(1 + \texttt{MNCAAS}_i) + \beta_2 \log(1 + \texttt{MNPRBA}_i) + \sum_{j \in \mathcal{J}} \alpha_j \, \texttt{SEGURO}_{ij} + \sum_{k \in \mathcal{K}} \kappa_k \, \texttt{INTERMEDIARIO}_{ik} + \sum_{m \in \mathcal{M}} \lambda_m \, \texttt{ZONA}_{im}\Big). \tag{6}$$

- **Modelo 2:** Extiende el anterior agregando características individuales: `EDAD`, $\texttt{EDAD}^2$ y `SEXO`.

$$P(\texttt{FUGA}_i = 1 \mid X_i) = G\Big(\beta_0 + \beta_1 \log(1 + \texttt{MNCAAS}_i) + \beta_2 \log(1 + \texttt{MNPRBA}_i) + \gamma_1 \texttt{EDAD}_i + \gamma_2 \texttt{EDAD}_i^2 + \gamma_3 \texttt{SEXO}_i + \sum_{j \in \mathcal{J}} \alpha_j \, \texttt{SEGURO}_{ij} + \sum_{k \in \mathcal{K}} \kappa_k \, \texttt{INTERMEDIARIO}_{ik} + \sum_{m \in \mathcal{M}} \lambda_m \, \texttt{ZONA}_{im}\Big). \tag{7}$$

[3]En las tablas se rotulan `EDAD` y $\texttt{EDAD}^2$, pero en la estimación se usa $\texttt{EDAD}_c = \texttt{EDAD} - \overline{\texttt{EDAD}}$.

- **Modelo 3:** Integra además indicadores de la relación con la aseguradora: DOC_IMPAGOS, ANTIGUEDAD_MESES, BENEFICIOS_RUT y Nueva_MNPMP.

$$
\begin{aligned}
P(\texttt{FUGA}_i = 1 \mid X_i) = G\Big(&\beta_0 + \beta_1 \log(1 + \texttt{MNCAAS}_i) + \beta_2 \log(1 + \texttt{MNPRBA}_i) \\
&+ \gamma_1 \texttt{EDAD}_i + \gamma_2 \texttt{EDAD}_i^2 + \gamma_3 \texttt{SEXO}_i \\
&+ \delta_1 \texttt{DOC\_IMPAGOS}_i + \delta_2 \texttt{ANTIGUEDAD\_MESES}_i \\
&+ \delta_3 \texttt{BENEFICIOS\_RUT}_i + \delta_4 \texttt{Nueva\_MNPMP}_i \\
&+ \sum_{j \in \mathcal{J}} \alpha_j \, \texttt{SEGURO}_{ij} + \sum_{k \in \mathcal{K}} \kappa_k \, \texttt{INTERMEDIARIO}_{ik} + \sum_{m \in \mathcal{M}} \lambda_m \, \texttt{ZONA}_{im} \Big).
\end{aligned}
\tag{8}
$$

- **Modelo 4:** Considera la incorporación de interacciones de dos vías entre variables explicativas, con el fin de capturar posibles efectos combinados.

$$
\begin{aligned}
P(\texttt{FUGA}_i = 1 \mid X_i) = G\Big(&\beta_0 + \beta_1 \log(1 + \texttt{MNCAAS}_i) + \beta_2 \log(1 + \texttt{MNPRBA}_i) \\
&+ \gamma_1 \texttt{EDAD}_i + \gamma_2 \texttt{EDAD}_i^2 + \gamma_3 \texttt{SEXO}_i \\
&+ \delta_1 \texttt{DOC\_IMPAGOS}_i + \delta_2 \texttt{ANTIGUEDAD\_MESES}_i \\
&+ \delta_3 \texttt{BENEFICIOS\_RUT}_i + \delta_4 \texttt{Nueva\_MNPMP}_i \\
&+ \theta_1 \big(\texttt{ANTIGUEDAD\_MESES}_i \times \texttt{Nueva\_MNPMP}_i\big) \\
&+ \sum_{j \in \mathcal{J}} \alpha_j \, \texttt{SEGURO}_{ij} + \sum_{k \in \mathcal{K}} \kappa_k \, \texttt{INTERMEDIARIO}_{ik} + \sum_{m \in \mathcal{M}} \lambda_m \, \texttt{ZONA}_{im} \Big).
\end{aligned}
\tag{9}
$$

En particular, se analizaron combinaciones entre variables continuas centrales ($\log(\texttt{MNPRBA})$, ANTIGUEDAD_MESES, DOC_IMPAGOS, BENEFICIOS_RUT, Nueva_MNPMP), interacciones entre variables demográficas (EDAD, SEXO) y características de la póliza, así como cruces de dichas variables con dummies representativos de zona geográfica (ZONA_$\star$) y

canal de venta (`INTERMEDIARIO_*`). Cada especificación candidata se estimó sobre la misma base que el Modelo 3 y fue evaluada mediante los criterios de información (AIC y BIC), junto con pruebas de razón de verosimilitud (LR) (ver Anexo A: Notas metodológicas). Los detalles comparativos y la selección de la especificación final se presentan en la sección de Resultados.

5.4. Enfoque de redes neuronales (MLP)

Como complemento a los modelos logit, se incorpora un clasificador neuronal del tipo *perceptrón multicapa* (MLP). Este tipo de modelo, perteneciente al campo del *deep learning*, permite capturar relaciones no lineales y complejas entre las variables explicativas, lo que abre la posibilidad de mejorar la capacidad predictiva respecto de un modelo estrictamente lineal. La variable dependiente sigue siendo la fuga de clientes (`FUGA=1`), por lo que la salida final de la red entrega una probabilidad estimada de fuga.

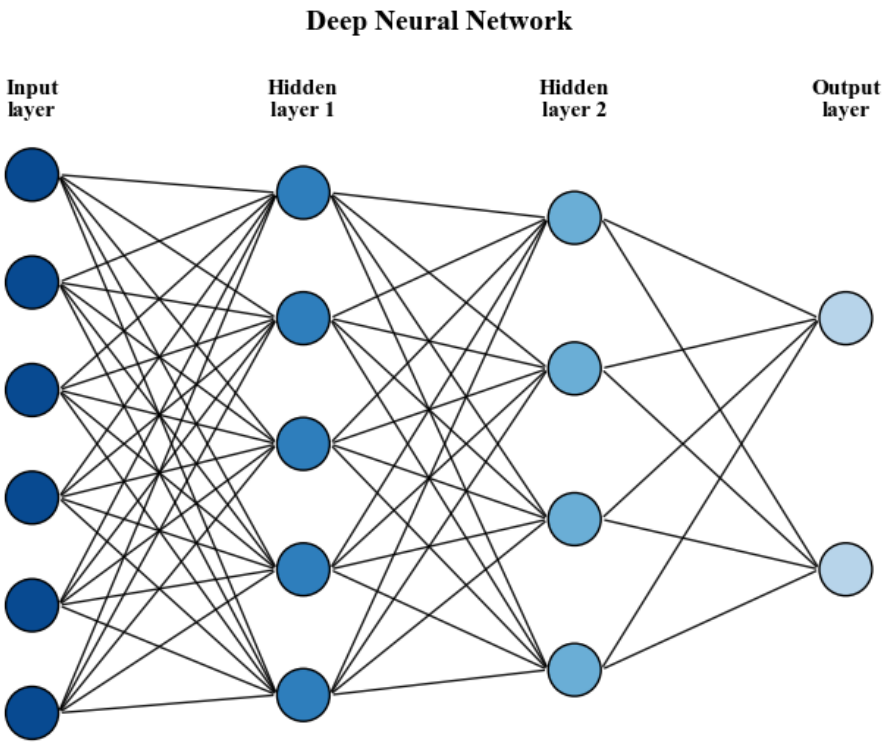


Figura 1: Esquema representativo del perceptrón multicapa (MLP) utilizado en el estudio, con una única salida sigmoide que entrega la probabilidad de fuga.

Arquitectura y funciones de activación. Una red neuronal está compuesta por **neuronas**, que son unidades de cálculo muy simples: cada neurona toma varias entradas $(x_1, x_2, \ldots, x_p)$, las combina linealmente mediante **pesos** $(w_1, w_2, \ldots, w_p)$ y un **sesgo** b, y luego aplica una **función de activación** $\phi(\cdot)$ que introduce no linealidad. Formalmente:

$$z = \sum_{j=1}^{p} w_j x_j + b, \qquad h = \phi(z).$$

Los **pesos** determinan cuánto "importa" cada variable de entrada en la predicción, mientras que el **sesgo** permite desplazar la función de activación hacia arriba o hacia abajo. Al entrenar la red, estos parámetros se van ajustando automáticamente para minimizar la función de pérdida.

Las neuronas se agrupan en **capas**. Una capa oculta recibe como entradas las salidas de la capa anterior y aplica este mismo proceso en paralelo con varias neuronas. El conjunto de capas define la arquitectura de la red. En este trabajo se consideran tres configuraciones de creciente complejidad:

- **Una capa oculta**: 64 neuronas.
- **Dos capas ocultas**: 128 y 64 neuronas, respectivamente.
- **Tres capas ocultas**: 128, 64 y 32 neuronas.

A mayor número de capas y neuronas, la red gana capacidad para capturar relaciones no lineales y complejas entre las variables de entrada, aunque también aumenta el riesgo de sobreajuste.

En cada capa oculta se evalúan dos funciones de activación estándar:

$$\text{ReLU}(u) = \text{máx}\{0, u\}, \tag{10}$$

que activa solo valores positivos y deja en cero los negativos, y

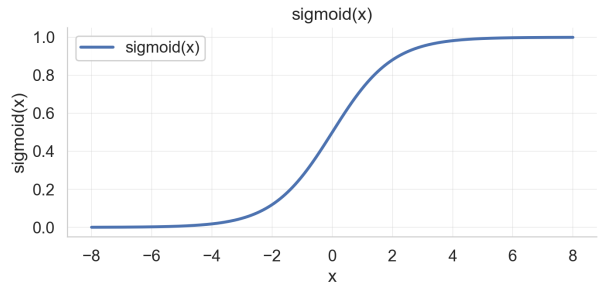
$$\tanh(u) = \frac{e^{u} - e^{-u}}{e^{u} + e^{-u}}, \tag{11}$$

que transforma la entrada en valores suaves entre -1 y 1. Ambas funciones permiten que la red modele patrones más complejos que los que podría capturar una regresión lineal.

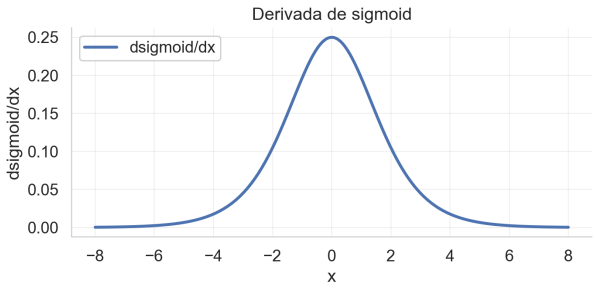
Ahora, la **capa de salida** utiliza una función sigmoide:

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

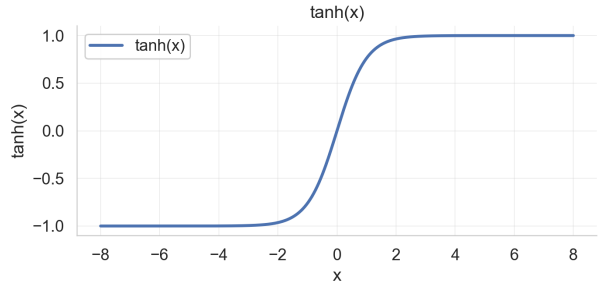
la cual devuelve un valor entre 0 y 1 que se interpreta directamente como la probabilidad de fuga de un cliente.



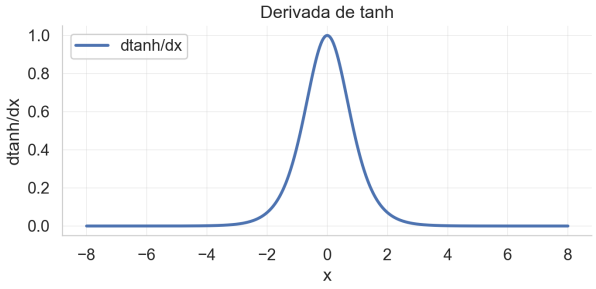
(a) Sigmoide



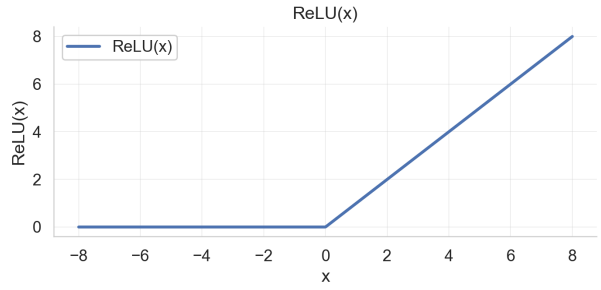
(b) Derivada de sigmoide



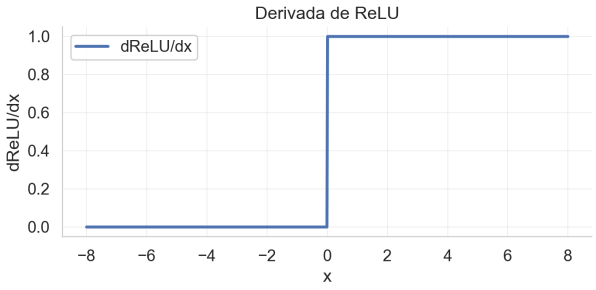
(c) Tanh



(d) Derivada de tanh



(e) ReLU



(f) Derivada de ReLU

Figura 2: Funciones de activación más utilizadas en redes neuronales (sigmoide, tanh y ReLU) y sus derivadas.

Función de pérdida y regularización. El entrenamiento de la red busca minimizar la entropía cruzada binaria (o *log-loss*), que mide qué tan bien las probabilidades estimadas por la red se

ajustan a las etiquetas reales (0 = no fuga, 1 = fuga). La expresión formal es:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{n}\sum_{i=1}^{n}\Big[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)\Big], \tag{12}$$

donde y_i es la etiqueta real (0 o 1) y $\hat{p}_i$ es la probabilidad estimada por el modelo de que ocurra fuga.

En términos prácticos:

- Si el cliente realmente se fuga ($y_i = 1$) y la red predice una probabilidad alta de fuga ($\hat{p}_i \approx 0{,}9$), la pérdida es baja:

$$-\log(0{,}9) \approx 0{,}105.$$

- Si el cliente se fuga ($y_i = 1$) pero la red asigna una probabilidad muy baja ($\hat{p}_i \approx 0{,}1$), la pérdida es alta:

$$-\log(0{,}1) \approx 2{,}303.$$

- De forma análoga, si el cliente no se fuga ($y_i = 0$), se penaliza fuertemente cuando el modelo predice con alta seguridad que sí habrá fuga ($\hat{p}_i$ cercano a 1).

De esta manera, la entropía cruzada no solo considera si la clasificación fue correcta, sino también qué tan confiado estaba el modelo en su predicción. Modelos que aciertan pero con baja probabilidad son penalizados más que aquellos que aciertan con alta seguridad.

Adicionalmente, se incluye una **penalización L2** sobre los pesos para evitar sobreajuste:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{CE}}(\theta) + \alpha\|\theta\|_2^2, \tag{13}$$

donde θ representa los parámetros de la red y $\alpha > 0$ es el hiperparámetro que controla la magnitud de la penalización. Este término actúa como un "freno": valores grandes de α generan modelos más simples y estables (aunque con riesgo de subajuste), mientras que valores muy pequeños permiten mayor flexibilidad (aunque con riesgo de memorizar los datos).

Preprocesamiento de variables. Antes del entrenamiento, todas las covariables continuas se estandarizan (media cero y desviación estándar uno) para facilitar la convergencia. En el caso de variables indicadoras (dummies), se evita el centrado de columnas muy dispersas para no distorsionar su significado. A diferencia del logit, no se agrega constante explícita, ya que cada capa posee términos de sesgo que cumplen ese rol.

Entrenamiento y optimización. El proceso de aprendizaje en una red neuronal consiste en ajustar los pesos y sesgos de cada neurona para que las probabilidades estimadas se acerquen lo más posible a los valores reales. Esto se logra mediante el algoritmo de retropropagación del gradiente: primero se calcula la pérdida en la salida (entropía cruzada), luego se obtienen los gradientes de dicha pérdida respecto de cada peso, y finalmente se actualizan los parámetros en la dirección que reduce el error.

El método general de actualización se basa en descenso por gradiente, que mueve los parámetros en pasos sucesivos hacia el mínimo de la función de pérdida. Dado que trabajar con todo el conjunto de datos en cada paso sería muy costoso, se emplean variantes estocásticas que usan subconjuntos (lotes o *batches*) para aproximar el gradiente.

El optimizador principal que se utiliza es **Adam**, que mejora al descenso estocástico clásico ajustando de manera adaptativa la tasa de aprendizaje de cada parámetro e incorporando un

término de momento que estabiliza la trayectoria de descenso. Esto lo hace especialmente robusto en problemas de alta dimensión. Como contraste, también se prueba **SGD** (descenso estocástico clásico), que actualiza directamente en la dirección del gradiente promedio del lote sin ajustes adaptativos.

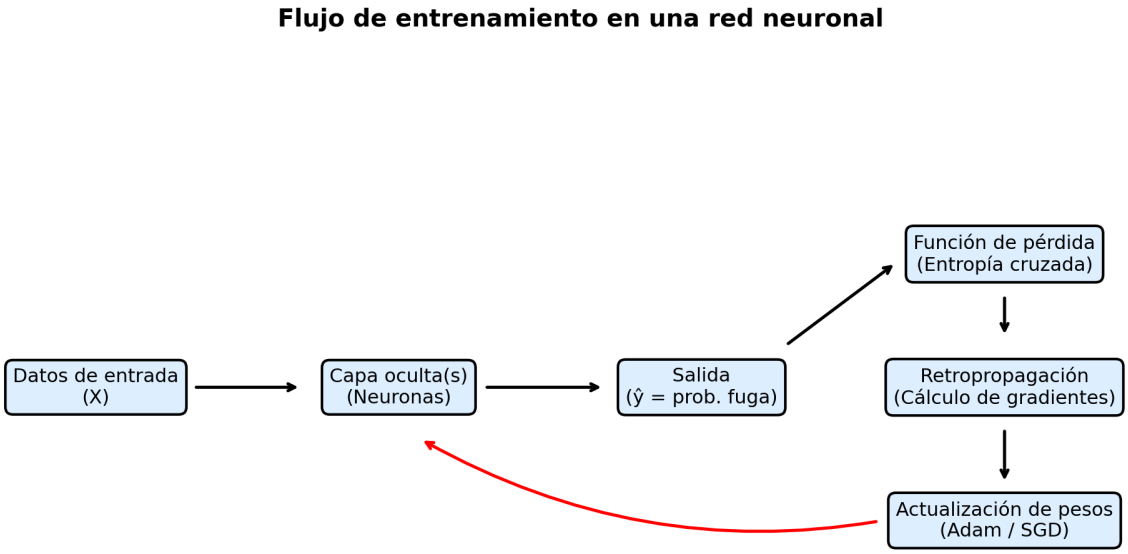


Figura 3: Flujo de entrenamiento en una red neuronal. El proceso parte con los datos de entrada, pasa por las capas ocultas y genera una probabilidad de fuga como salida. A partir de ella se calcula la función de pérdida (entropía cruzada), se realiza retropropagación para obtener gradientes, y se actualizan los pesos mediante optimizadores como Adam o SGD.

Entre los hiperparámetros más relevantes destacan:

- **Tasa de aprendizaje inicial** (η_0): define el tamaño de cada paso de actualización de los pesos. Una tasa demasiado alta puede hacer que el algoritmo oscile sin converger, mientras que una demasiado baja conduce a un entrenamiento muy lento. En este trabajo se exploran valores en el rango $[10^{-4},\, 5 \cdot 10^{-2}]$.

- **Tamaño de lote (*batch*)**: número de observaciones procesadas antes de realizar una actualización de pesos. Lotes pequeños (p. ej. 128) generan gradientes más ruidosos que pueden favorecer la exploración de mínimos, mientras que lotes grandes (p. ej. 1024) producen gradientes más estables y una convergencia más suave.

- **Regularización L2 (α)**: penaliza pesos excesivamente grandes para evitar que la red memorice la muestra de entrenamiento. Valores altos de α producen modelos más simples pero con riesgo de subajuste, mientras que valores muy bajos permiten mayor flexibilidad con riesgo de sobreajuste. En la búsqueda se considera $\alpha \in [10^{-5},\, 10^{-2}]$.

Finalmente, para prevenir el sobreajuste se implementa la técnica de detención temprana (*early stopping*). Durante el entrenamiento, una fracción del conjunto de entrenamiento se utiliza como validación: si tras varias iteraciones el desempeño en validación deja de mejorar, el proceso se interrumpe automáticamente. Esto asegura que el modelo retenga una buena capacidad de generalización y no se ajuste en exceso a los datos de entrenamiento.

Partición de datos y validación. Los datos se dividen de forma estratificada en 70 % entrenamiento y 30 % prueba, preservando la proporción de fugas. En el conjunto de entrenamiento se ejecuta una búsqueda aleatoria de hiperparámetros con validación cruzada de tres pliegues, utilizando como métrica de selección el AUC promedio. Además, se deben explorar combinaciones de funciones de activación (`relu`, `tanh`), optimizadores (`adam`, `sgd`), tasas de regularización (α entre 10^{-5} y 10^{-2}), tasas de aprendizaje iniciales (entre 10^{-4} y $5 \cdot 10^{-2}$) y tamaños de lote (128, 256, 512, 1024). De esta manera, al encontrarse el mejor conjunto de hiperparámetros, se reentrena en el 70 % de los datos y luego se evalúa en el 30 % de prueba.

5.5. Evaluación predictiva: matriz de confusión, umbrales y métricas

Probabilidades y regla de clasificación. Una vez estimado el modelo logit, se obtienen para cada observación i las probabilidades ajustadas $\hat{p}_i = \Lambda(x_i'\hat{\beta})$, donde $\Lambda(\cdot)$ es la CDF logística. Dado que la variable de interés es binaria ($\texttt{FUGA} = 1$ si la póliza se anula/no renueva; 0 en caso contrario), se define una regla de decisión que asigna la clase predicha como

$$\hat{y}_i(t) = 1\{\hat{p}_i \geq t\},$$

donde $t \in (0, 1)$ es el *umbral* de clasificación.

Criterios de umbral considerados. Se emplean tres criterios estándar y complementarios:

1. **Umbral** $0{,}5$: regla convencional que clasifica como "fuga" a quienes tienen $\hat{p}_i \geq 0{,}5$.

2. **Índice de Youden**: se elige el umbral t^* que maximiza $J(t) = \text{TPR}(t) - \text{FPR}(t)$ sobre la curva ROC. Aquí TPR (tasa de verdaderos positivos o sensibilidad) mide la proporción de fugas correctamente identificadas sobre el total de fugas reales, mientras que FPR (tasa de falsos positivos) mide la proporción de clientes sin fuga que fueron clasificados erróneamente como fuga. El índice de Youden busca el punto de corte que mejor equilibre ambos aspectos.

3. **Umbral que maximiza F1**: se selecciona t para el cual la medida F1 alcanza su mayor valor. La medida F1 combina en un único indicador la *precision* (qué fracción de los clasificados como fuga efectivamente lo son) y el *recall* (qué fracción de las fugas reales se logra detectar).

Matriz de confusión y métricas derivadas. La matriz de confusión permite organizar los aciertos y errores de un clasificador binario en cuatro categorías:

- **Verdaderos Positivos (TP)**: clientes que efectivamente se fugaron y fueron correctamente clasificados como fuga.
- **Falsos Positivos (FP)**: clientes que no se fugaron, pero el modelo predijo erróneamente como fuga.
- **Verdaderos Negativos (TN)**: clientes que no se fugaron y fueron correctamente clasificados como no fuga.
- **Falsos Negativos (FN)**: clientes que sí se fugaron, pero el modelo predijo erróneamente como no fuga.

A partir de estos conteos se derivan las principales métricas de desempeño utilizadas en este trabajo:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{14}$$

Mide la proporción total de clasificaciones correctas (fugas y no fugas) sobre el total de observaciones. Es un indicador global, aunque puede ser engañoso en presencia de clases desbalanceadas.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

También llamada valor predictivo positivo. Indica qué fracción de los clientes clasificados como "fuga" lo son realmente. Una precisión alta significa que el modelo comete pocos falsos positivos.

$$\text{Recall (Sensibilidad)} = \frac{TP}{TP + FN} \tag{16}$$

También conocida como tasa de verdaderos positivos (TPR). Mide la proporción de fugas reales que fueron correctamente identificadas. Una sensibilidad alta significa que se pierden pocos casos de fuga (pocos falsos negativos).

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{17}$$

Es la media armónica entre precisión y sensibilidad. Resume ambas dimensiones en un solo número, siendo especialmente útil cuando se busca un balance entre detectar la mayor cantidad de fugas y mantener un nivel de falsos positivos aceptable.

$$\text{AUC} = \int_0^1 TPR(FPR)\, d(FPR) \tag{18}$$

El área bajo la curva ROC (AUC) mide la capacidad discriminante del modelo en todo el rango de posibles umbrales. Un AUC cercano a 1 indica un excelente poder de separación entre clientes que se fugan y los que no; un AUC de 0,5 refleja un desempeño equivalente a asignar las clases al azar.

Implementación

En este trabajo se define como positivo de referencia la condición `FUGA=1`. Para cada especificación estimada (M1, M2 y M3) se generan las probabilidades individuales de fuga y, a partir de ellas, se aplican los tres criterios de umbral descritos (0,5; Youden; F1). Con cada punto de corte se construye la matriz de confusión correspondiente y se calculan las métricas derivadas

(accuracy, precision, recall, F1).

De forma complementaria, el área bajo la curva ROC (AUC) se calcula directamente sobre el vector de probabilidades ajustadas, ya que no depende de un umbral específico. De este modo, el protocolo seguido separa con claridad dos aspectos distintos: por un lado, la elección del umbral como decisión operativa de clasificación; y, por otro, la medición de la capacidad discriminante intrínseca del modelo a través del AUC. Esto permite evaluar de manera integral tanto la calidad de la clasificación puntual como el poder predictivo global de cada especificación.

6. Resultados

En esta sección se presentan los resultados obtenidos a partir de las regresiones logísticas estimadas conforme a las especificaciones descritas previamente.

Los modelos se estimaron mediante el método de máxima verosimilitud, y los asteriscos que acompañan a los coeficientes indican su nivel de significancia estadística: ***p-valor<0,01**, **p-valor<0,05** y *p-valor<0,10**. Estos niveles permiten evaluar la solidez estadística de los efectos estimados sobre la probabilidad de fuga.

Tabla 3: Coeficientes estimados en modelos logit

Variable	**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**
$\log(1 + \texttt{MNCAAS})$	0.3535***	0.2374***	0.1158***	0.0679***
$\log(1 + \texttt{MNPRBA})$	-1.1949***	-0.8796***	-0.9708***	-0.8165***
EDAD		-0.0248***	-0.0264***	-0.0271***
$\texttt{EDAD}^2$		0.0008***	0.0006***	0.0006***
SEXO (0:masc./1:fem.)		-0.0096	-0.0300***	-0.0292***
DOC_IMPAGOS			0.0457***	0.0486***
ANTIGUEDAD_MESES			-0.0357***	-0.0359***
BENEFICIOS_RUT			-0.0087***	-0.0089***
Nueva_MNPMP			0.0572***	-0.0492***
ANTIGUEDAD_MESES × Nueva_MNPMP				-0.0082***
Controles: SEGURO_⋆, INTERMEDIARIO_⋆, ZONA_⋆	Sí	Sí	Sí	Sí

Notas: *** p-valor<0,01; ** p-valor<0,05; * p-valor<0,10.

Patrones constantes. En todas las especificaciones, $\log(1 + \texttt{MNCAAS})$ se asocia positivamente con la probabilidad de fuga y $\log(1 + \texttt{MNPRBA})$ lo hace negativamente. Al añadir regresoras (M2 y M3) la magnitud de ambos efectos se reduce en valor absoluto, como es esperable al controlar por mayor heterogeneidad (coeficientes: MNCAAS $0.3535 \rightarrow 0.1158$; MNPRBA $-1.1949 \rightarrow -0.9708$).

Edad y no linealidad. En M2 y M3, EDAD es negativa y $\texttt{EDAD}^2$ positiva, lo que sugiere una relación en "U": la probabilidad de fuga disminuye en edades medias y aumenta en edades altas. El punto de mínima probabilidad se ubica por encima de la media muestral en aproximadamente 15,7 años (M2) y 20,7 años (M3).[4]

Tabla 4: Efectos marginales promedio (AME) en modelos logit

Variable	Modelo 1	Modelo 2	Modelo 3	Modelo 4
$\log(1+\texttt{MNCAAS})$	0.0699***	0.0464***	0.0289***	0.0170***
$\log(1+\texttt{MNPRBA})$	-0.2363***	-0.1719***	-0.2425***	-0.2040***
EDAD		-0.0048***	-0.0066***	-0.0068***
$\texttt{EDAD}^2$		0.0002***	0.0002***	0.0002***
SEXO (0:masc./1:fem.)		-0.0019	-0.0075***	-0.0073***
DOC_IMPAGOS			0.0114***	0.0121***
ANTIGUEDAD_MESES			-0.0089***	-0.0090***
BENEFICIOS_RUT			-0.0022***	-0.0022***
Nueva_MNPMP			0.0143***	-0.0123***
ANTIGUEDAD_MESES × Nueva_MNPMP				-0.0020***
Controles: SEGURO_*, INTERMEDIARIO_*, ZONA_*	Sí	Sí	Sí	Sí

Notas: *** p-valor<0,01; ** p-valor<0,05; * p-valor<0,10.

Los AME permiten una interpretación directa en probabilidades. Un incremento en $\log(1+\texttt{MNCAAS})$ se asocia con aumentos entre 0,07 y 0,017 puntos porcentuales en la probabilidad de fuga (M1–M4). En contraste, mayores valores de $\log(1+\texttt{MNPRBA})$ reducen la probabilidad de fuga en torno a 0,24 puntos porcentuales (M3).

Relación con la aseguradora (M3). DOC_IMPAGOS incrementa la probabilidad de fuga (AME ≈ 0.0114 por unidad; si fuese binaria, ~ 1.14 pp al pasar de 0 a 1). ANTIGUEDAD_MESES la reduce (AME ≈ -0.0089 por mes; ~ -10.7 pp en 12 meses, ceteris paribus). BENEFICIOS_RUT

[4]Se usa $\text{EDAD}_c = \text{EDAD} - \overline{\text{EDAD}}$. En un logit con cuadrático, el mínimo es $\text{EDAD}^* = \overline{\text{EDAD}} - \beta_1/(2\beta_2)$. Con los coeficientes estimados: $-\beta_1/(2\beta_2) \approx 15.7$ (M2) y 20.7 (M3).

también la reduce (AME ≈ -0.0022; si es 0/1, ~ -0.22 pp). `Nueva_MNPMP` presenta efecto positivo (AME ≈ 0.0143; si es 0/1, ~ 1.43 pp). La mayoría son estadísticamente significativas; por ejemplo, `SEXO` no es significativo en M2 (p≈ 0.33), pero sí en M3 (p< 0.01).

Bondad de ajuste incremental. Ampliar la especificación mejora significativamente el ajuste:

- M1→M2: LR $= 3291{,}24$, $df = 3$, $p < 0{,}001$.
- M2→M3: LR $= 59043{,}47$, $df = 4$, $p < 0{,}001$.

De esta manera, los tests de Wald por bloques confirman la relevancia conjunta de las variables añadidas. Para M1→M2 se prueba $H_0 : \{\beta_{\texttt{EDAD}}, \beta_{\texttt{EDAD}^2}, \beta_{\texttt{SEXO}}\} = 0$ y se rechaza (Wald $\chi^2(3) = 3210{,}45$, $p < 0{,}001$).

Para M2→M3 se prueba $H_0 : \{\beta_{\texttt{DOC\_IMPAGOS}}, \beta_{\texttt{ANTIGUEDAD\_MESES}}, \beta_{\texttt{BENEFICIOS\_RUT}}, \beta_{\texttt{Nueva\_MNPMP}}\} = 0$ y también se rechaza (Wald $\chi^2(4) = 44551{,}19$, $p < 0{,}001$).

Identificación. La matriz de diseño presenta rango completo en todas las especificaciones ($\text{rank}(X_1) = 73/73$, $\text{rank}(X_2) = 76/76$, $\text{rank}(X_3) = 80/80$), lo que descarta colinealidad perfecta y garantiza identificación y errores estándar finitos.[5]

Interacciones (M4) y lectura. Por otro lado, al extender el modelo con términos de interacción se identificó como relevante el cruce entre `ANTIGUEDAD_MESES` y `Nueva_MNPMP`. El coeficiente es negativo (-0.0082, $p < 0.01$) y el AME es -0.0020, indicando que el efecto reductor de la antigüedad sobre la probabilidad de fuga es más pronunciado en clientes con niveles altos de prima adicional. Frente a M3, se observan mejoras importantes de ajuste ($\Delta AIC = -626{,}7$;

[5]Esto no descarta colinealidad alta, pero sí la perfecta; los resultados no muestran inestabilidades atribuibles a ello.

$\Delta BIC = -616{,}1$) y un test de razón de verosimilitud altamente significativo (LR $= 628{,}7$, $df = 1$, $p < 0{,}001$). Los tests de Wald confirman que la interacción es globalmente distinta de cero.

Interpretación en los datos. El signo negativo de la interacción confirma que la relación entre antigüedad y probabilidad de fuga depende del nivel de prima adicional: a mayor `Nueva_MNPMP`, más fuerte es el efecto protector de la antigüedad. En cambio, cuando `Nueva_MNPMP` está bajo la media, el efecto de la antigüedad sobre la retención se atenúa. Esto sugiere que la permanencia no depende solo de la duración de la relación contractual o del monto de prima, sino de la combinación de ambas dimensiones.

6.1. Matriz de confusión y desempeño predictivo

Esta sección sintetiza el desempeño predictivo de los modelos logit mediante sus matrices de confusión y métricas asociadas. El foco es evaluar cómo varía la capacidad de clasificación al modificar el umbral de probabilidad de fuga. Para ello, se efectuó un barrido de umbrales en $[0, 1]$ y se reportan tres puntos de referencia:

- **Umbral estándar ($t = 0{,}5$):** punto de corte convencional que asume costos de error similares entre clases.
- **Umbral óptimo de Youden:** maximiza el índice $J = \mathrm{TPR} - \mathrm{FPR}$ (balance sensibilidad–especificidad).
- **Umbral que maximiza F_1:** optimiza F_1, que combina *precision* y *recall*.

Dado que en este problema el costo de *no detectar* a un cliente en riesgo (falso negativo) es mayor que el de intervenir sobre un cliente que se queda (falso positivo), la métrica F_1 resulta

especialmente apropiada. A continuación se presentan las tablas y, tras cada una, su lectura.

Tabla 5: Matrices de confusión con umbral fijo 0,5

Modelo 1		
	No fuga	Fuga
Real: No fuga	69,189	61,949
Real: Fuga	30,667	139,179
Modelo 2		
	No fuga	Fuga
Real: No fuga	68,045	63,093
Real: Fuga	27,935	141,911
Modelo 3		
	No fuga	Fuga
Real: No fuga	90,142	40,996
Real: Fuga	28,187	141,659

Notas: Filas = valores reales; columnas = valores predichos. Los encabezados dentro de cada bloque indican la clase estimada.

En la Tabla 5 con el umbral de $0{,}5$, los tres modelos muestran un desempeño correcto. Destaca **M3** (exactitud $0{,}770$, precisión $0{,}776$, sensibilidad $0{,}834$), lo que anticipa una mejor discriminación que M1 y M2. Aun así, el volumen de falsos negativos sugiere ajustar el punto de corte para privilegiar la detección de fugas.

Tabla 6: Matrices de confusión con umbral óptimo de Youden (M1: 0,581; M2: 0,550; M3: 0,580)

Modelo 1		
	No fuga	Fuga
Real: No fuga	81,583	49,555
Real: Fuga	44,522	125,324
Modelo 2		
	No fuga	Fuga
Real: No fuga	77,821	53,317
Real: Fuga	38,956	130,890
Modelo 3		
	No fuga	Fuga
Real: No fuga	99,450	31,688
Real: Fuga	38,427	131,419

Nota: El umbral de Youden se determina maximizando $J = \mathrm{TPR} - \mathrm{FPR}$ sobre la curva ROC.

En cuanto al criterio de Youden (Tabla 6), se tiene que aumenta la *precision* y reduce falsos positivos, a cambio de sacrificar algo de *recall*. Para **M3** ($t_Y \approx 0{,}58$) la precisión asciende a $0{,}806$ y la sensibilidad queda en $0{,}774$ (exactitud $0{,}767$), adecuado cuando se busca un balance entre capturar fugas y evitar intervenciones innecesarias.

Tabla 7: Matrices de confusión con umbral que maximiza F1 (M1: 0,365; M2: 0,406; M3: 0,373)

Modelo 1		
	No fuga	Fuga
Real: No fuga	36,108	95,030
Real: Fuga	5,670	164,176
Modelo 2		
	No fuga	Fuga
Real: No fuga	50,287	80,851
Real: Fuga	13,249	156,597
Modelo 3		
	No fuga	Fuga
Real: No fuga	74,386	56,752
Real: Fuga	15,685	154,161

Nota: El umbral se determinó mediante un barrido exhaustivo, eligiendo el valor que maximiza la medida F_1.

Al optimizar F_1 se prioriza la recuperación de fugas. En **M3**, el punto $t_{F_1} = 0{,}373$ logra sensibilidad $0{,}908$ y precisión $0{,}731$, es decir, una captura sustantiva de casos a costa de un aumento controlado de falsos positivos; esto es consistente con los objetivos operativos del problema.

Tabla 8: Métricas globales por modelo y criterio de umbral

Modelo	Umbral	Accuracy	Precision	Recall	F1	AUC
M1	0,50	0,692	0,692	0,819	0,750	0,741
M2	0,50	0,698	0,692	0,836	0,757	0,748
M3	0,50	0,770	0,776	0,834	0,804	0,850
M1	Youden	0,687	0,717	0,738	0,727	0,741
M2	Youden	0,693	0,711	0,771	0,739	0,748
M3	Youden	0,767	0,806	0,774	0,789	0,850
M1	Mejor F1	0,665	0,633	0,967	0,765	0,741
M2	Mejor F1	0,687	0,660	0,922	0,769	0,748
M3	Mejor F1	0,759	0,731	0,908	0,810	0,850

Lectura: Precision es el valor predictivo positivo y *Recall* la sensibilidad. El AUC se calcula sobre probabilidades (independiente del umbral).

La comparación consolidada confirma la superioridad de **M3** en todas las métricas globales; su AUC $= 0{,}850$ respalda su capacidad discriminante sin depender de un corte particular. La Figura 4 resume la evolución de métricas con el umbral y muestra el máximo de F_1 en torno a $t_{F_1} = 0{,}373$.

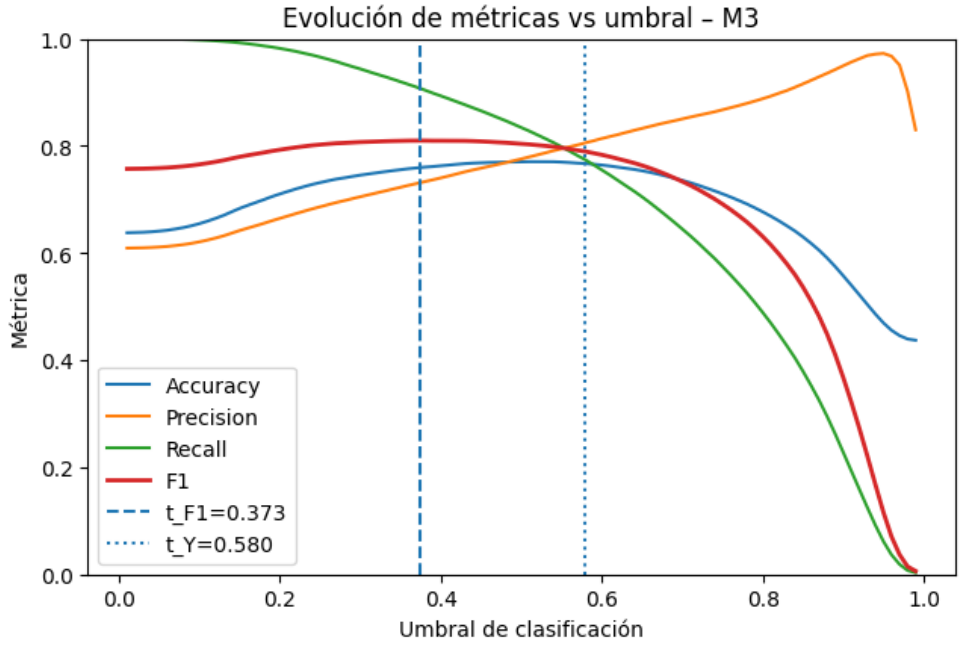


Figura 4: Evolución de métricas en función del umbral de clasificación para el Modelo 3.

En síntesis, **M3** es la especificación preferente. Operacionalmente, se propone usar el umbral que maximiza F_1 ($t = 0{,}373$) para una detección amplia de riesgo de fuga.

6.2. Desempeño predictivo con redes neuronales

Tras evaluar los logit, se contrasta su desempeño con redes neuronales multicapa (*Multi-Layer Perceptron*, MLP) para verificar si la mayor flexibilidad no lineal mejora la capacidad predictiva. Se entrenaron tres arquitecturas: una capa oculta (64), dos capas ocultas (128, 64) y tres capas ocultas (128, 64, 32), usando un 30 % de prueba estratificado. Como antes, se reportan métricas en $t = 0{,}5$, en el umbral de Youden y en el umbral que maximiza F_1; además, se realizó el barrido para identificar explícitamente el t que maximiza F_1 en cada arquitectura.

Tabla 9: Métricas en conjunto de prueba por arquitectura y criterio de umbral (MLP)

Modelo	Umbral	Accuracy	Precision	Recall	F1	AUC
MLP_1capa	0,50	0,899	0,917	0,902	0,910	0,962
MLP_1capa	Youden	0,899	0,926	0,892	0,908	0,962
MLP_1capa	Mejor F_1	0,899	0,914	0,905	0,910	0,962
MLP_2capas	0,50	0,904	0,925	0,904	0,914	**0,967**
MLP_2capas	Youden	0,902	0,940	0,883	0,910	**0,967**
MLP_2capas	Mejor F_1	0,904	0,908	0,923	**0,916**	**0,967**
MLP_3capas	0,50	0,904	0,931	0,896	0,913	0,967
MLP_3capas	Youden	0,904	0,931	0,896	0,913	0,967
MLP_3capas	Mejor F_1	0,905	0,919	0,911	0,915	0,967

Nota: Resultados en el 30 % de prueba. El sombreado destaca la arquitectura con mejor AUC (MLP_2capas).

Según lo que se muestra en la Tabla 9, las tres arquitecturas obtienen AUC entre $0{,}962$ y $0{,}967$, claramente por sobre el logit ($0{,}850$). Mientras que el modelo escogido es: **MLP_2capas (128, 64)**, ya que este logra el mayor AUC en prueba ($0{,}967$) y el mejor F_1 con su corte óptimo ($0{,}916$), por lo que se adopta como referencia. En términos de umbral, el mejor F_1 incrementa la sensibilidad respecto de $t = 0{,}5$ preservando un equilibrio global favorable.

Tabla 10: Arquitecturas y configuraciones base de los modelos MLP

Arquitectura	**Activación**	**Optimizador**	**Batch**
MLP_1capa (64)	ReLU	Adam	1024
MLP_2capas (128,64)	**tanh**	**Adam**	**512**
MLP_3capas (128,64,32)	ReLU	Adam	1024

La arquitectura de dos capas combina activación *tanh* con *Adam* y tamaño de lote intermedio (512), configuración que en esta base aporta estabilidad y buena generalización.

Tabla 11: Hiperparámetros ganadores y desempeño en prueba

Arquitectura	α (L2)	η_0 (LR inicial)	**AUC (test)**
MLP_1capa (64)	4,67e-3	2,58e-2	0,962
MLP_2capas (128,64)	**1,96e-4**	**2,61e-4**	**0,967**
MLP_3capas (128,64,32)	3,20e-4	3,42e-3	0,967

Notas: α es la penalización L2; η_0 es la tasa de aprendizaje inicial. En todos los casos se utilizó *early stopping*.

El mejor AUC en prueba se obtiene con **MLP_2capas**, usando regularización L2 baja ($\alpha \approx 1{,}96 \times 10^{-4}$) y tasa de aprendizaje inicial pequeña ($\eta_0 \approx 2{,}61 \times 10^{-4}$).

Tabla 12: Matrices de confusión: mejor arquitectura (MLP_2capas) bajo tres umbrales

MLP_2capas ($t = 0{,}50$)		
	No fuga	Fuga
Real: No fuga	35,597	3,745
Real: Fuga	4,907	46,047
MLP_2capas (Youden, $t \approx 0{,}595$**)**		
	No fuga	Fuga
Real: No fuga	36,463	2,879
Real: Fuga	5,966	44,988
MLP_2capas (Mejor F_1**,** $t \approx 0{,}409$**)**		
	No fuga	Fuga
Real: No fuga	34,594	4,748
Real: Fuga	3,931	47,023

Lectura: Filas = valores reales; columnas = valores predichos. El umbral de F_1 maximiza la captura de fugas (mayor *recall*); Youden eleva la *precision* al reducir falsos positivos.

Con t que maximiza F_1 ($\approx 0{,}409$), la red aumenta la sensibilidad y mantiene $F_1 = 0{,}916$.

El umbral de Youden ($\approx 0{,}595$) prioriza *precision* ($0{,}940$) reduciendo falsos positivos, mientras que $t = 0{,}50$ queda en un punto intermedio.

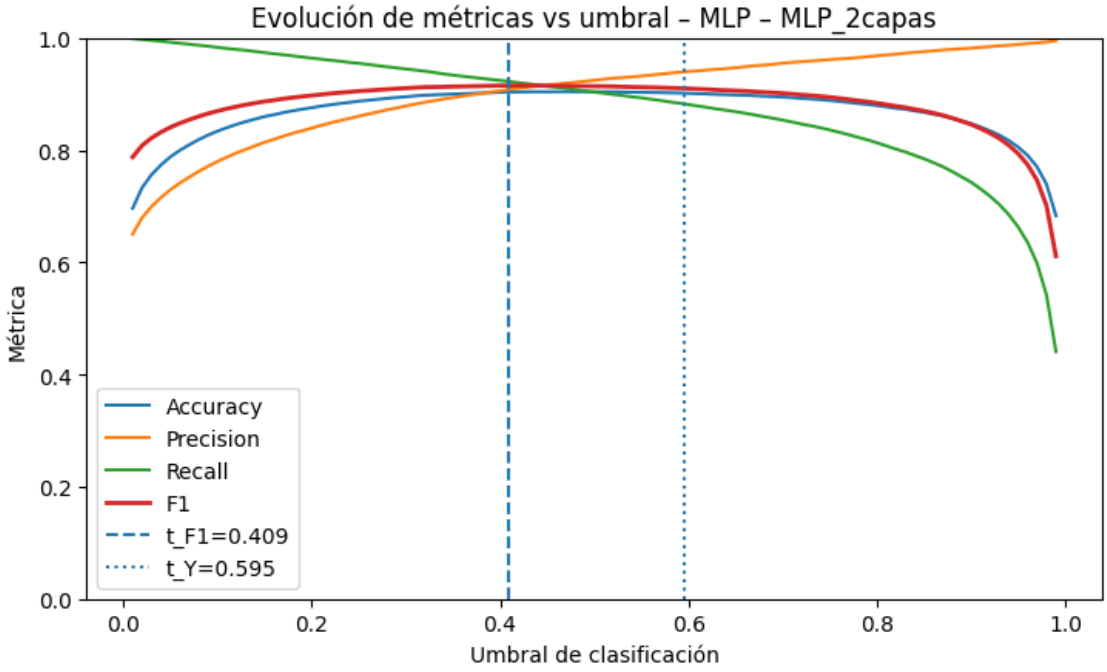


Figura 5: Evolución de métricas en función del umbral para la arquitectura MLP_2capas. Se indican los cortes $t_{F_1} \approx 0{,}409$ (línea discontinua) y $t_Y \approx 0{,}595$ (línea punteada).

En síntesis, las redes neuronales superan con holgura el desempeño del logit. El modelo **MLP_2capas (128, 64)** queda seleccionado como referencia (AUC $= 0{,}967$) y emplea el umbral que maximiza F_1 como punto de corte principal. Este criterio ofrece el mejor equilibrio entre precisión y sensibilidad, permitiendo una detección temprana y eficaz de clientes con alta probabilidad de fuga.

7. Discusión

7.1. Modelos logit y red neuronal

En este estudio se abordó la problemática de la fuga de clientes mediante dos enfoques complementarios: el modelo logit y una red neuronal artificial. Ambos presentaron rendimientos diferenciados que permiten extraer conclusiones relevantes sobre sus fortalezas y limitaciones, así como sobre el valor que aportan al análisis del fenómeno.

El modelo logit, si bien mostró un desempeño más limitado en términos predictivos, con métricas de exactitud y *recall* inferiores a las obtenidas por la red neuronal, entrega un valor esencial a través de su capacidad inferencial. Gracias a la estimación de coeficientes y su significancia estadística, este modelo permite identificar qué factores inciden en la probabilidad de fuga y en qué dirección lo hacen. Los resultados sugieren, por ejemplo, que la mayor antigüedad contractual disminuye la propensión a migrar, lo que es coherente con estudios previos que señalan que la permanencia de clientes en servicios financieros refuerza la lealtad debido a los costos hundidos percibidos y a la acumulación de confianza en la institución (Gupta et al., 2006). Asimismo, se aprecia un efecto diferenciado por género: las mujeres tienden a presentar una menor probabilidad de incumplimiento, hallazgo que concuerda con la evidencia en microfinanzas y banca que reporta un comportamiento de pago más responsable en este segmento (D'Espallier et al., 2011). De esta forma, aunque el logit no sea el mejor modelo en términos de predicción, sí ofrece un sustento explicativo fuerte para fundamentar políticas orientadas a la retención.

En contraste, la red neuronal logró un desempeño superior en las métricas predictivas, destacando particularmente en el *recall* y el F1-score. Esta capacidad la posiciona como una herramienta ideal para anticipar qué clientes se encuentran en riesgo de fuga, en línea con lo docu-

mentado por investigaciones recientes que muestran cómo los modelos de aprendizaje profundo superan a los estadísticos tradicionales en contextos de predicción de fuga (AbdelAziz et al., 2025; Mena et al., 2019; Bogaert and Delaere, 2023). La flexibilidad de las redes neuronales para capturar relaciones no lineales y patrones complejos entre variables constituye una ventaja evidente para generar sistemas de alerta temprana y segmentaciones de riesgo más precisas. Sin embargo, esta potencia predictiva tiene como contraparte su carácter de comportarse como una caja negra, lo cual dificulta la interpretación causal de los resultados. Este dilema entre capacidad predictiva e interpretabilidad ha sido ampliamente discutido en la literatura sobre modelado de fuga de clientes (Breiman, 2001; Ribeiro et al., 2016), y representa un desafío relevante en industrias reguladas como la aseguradora, donde la transparencia en la toma de decisiones es crucial.

En consecuencia, la comparación entre ambos enfoques muestra que ninguno, por sí solo, constituye una solución integral a la problemática, pues el modelo logit aporta claridad explicativa sobre los factores que impulsan la fuga, mientras que la red neuronal entrega mayor capacidad predictiva para anticipar escenarios futuros. Sin embargo, en conjunto, su aplicación complementaria permite diseñar políticas estratégicas de fidelización basadas en evidencia, a la vez que implementar herramientas operativas de alerta temprana. Esta perspectiva ”híbrida” ha sido planteada también en la literatura reciente, que recomienda combinar modelos interpretables con otros de alto desempeño predictivo para equilibrar transparencia y efectividad (Ahn et al., 2024; Castro et al., 2023).

7.2. Importancia en la literatura

Por otro lado, en cuanto a los vínculos que existen con la literatura, se tiene que los hallazgos de este trabajo contribuyen a la predicción de fuga de clientes (*customer churn*) de dos maneras principales: por una parte, al confirmar tendencias observadas globalmente respecto al desempeño relativo de modelos explicativos (como el logit) frente a modelos predictivos más complejos (como redes neuronales), y por otra, al aportar evidencia en el contexto latinoamericano de seguros, donde hay pocos estudios con énfasis tanto inferencial como predictivo.

Se tiene que a nivel global, numerosos estudios han mostrado que los modelos de aprendizaje automático y aprendizaje profundo tienden a superar a los modelos estadísticos tradicionales en tareas de predicción de fuga. Por ejemplo, (AbdelAziz et al., 2025) realizaron una evaluación comparativa de varios modelos ML y DL en los sectores de internet, telecomunicaciones y seguros, e identificaron que las redes neuronales y los modelos ensamblados (*ensemble*) mejoran las métricas de *recall*, precisión y AUC en comparación con la regresión logística o XGBoost. De manera similar, estudios como el de (Sharma, 2013) ya sugerían que las redes neuronales podían capturar patrones no lineales que modelos logit tradicionales no detectaban, especialmente cuando se dispone de variables de comportamiento de clientes. Otros trabajos que utilizan datos secuenciales (series temporales) han demostrado que los modelos basados en memoria a corto y largo plazo (cuya sigla en inglés corresponde a LSTM: Long Short-Term Memory) superan consistentemente a las regresiones logísticas al considerar variables dinámicas como recencia, frecuencia y montos transados (Mena et al., 2019).

En el contexto latinoamericano, un estudio destacado es el de (Henao Madrigal et al., 2020), que investiga los determinantes de fuga en clientes multiproducto en la industria aseguradora de

la región y desarrolla modelos predictivos aplicados a este entorno. Sus resultados muestran que variables socioeconómicas, geográficas y características específicas de los productos juegan un rol determinante en la fuga, lo que sugiere que factores locales (económicos, culturales o de competencia) modulan los resultados que en otros países suelen considerarse "clásicos". Este hallazgo confirma que, aunque los patrones generales sean similares, la magnitud y la significancia de ciertas variables pueden diferir entre mercados, lo que justifica que este estudio se enfoque en la aseguradora estudiada y tome en cuenta los atributos locales de sus datos.

Además, la combinación de ambos modelos en este trabajo se relaciona con un tema cada vez más discutido en la literatura: la compensación (trade-off) entre poder explicativo e intensidad predictiva. En otras palabras, los modelos que predicen muy bien suelen ser difíciles de interpretar, mientras que los modelos más claros en su explicación tienden a tener un menor desempeño en pronósticos. Esto es especialmente relevante en sectores como el asegurador, donde no solo importa anticipar quién podría abandonar, sino también entender por qué, ya que las decisiones están sujetas a costos importantes, regulaciones y consideraciones éticas. Frente a este dilema, varios autores proponen soluciones intermedias, como el uso de modelos híbridos que combinan algoritmos complejos de "caja negra" con métodos interpretables (por ejemplo, la regresión logística o técnicas de análisis *post-hoc*). De hecho, algunos estudios han mostrado aplicaciones donde las predicciones de redes neuronales sirven como insumos en modelos logísticos, o donde se aplican herramientas de interpretabilidad sobre modelos avanzados para obtener explicaciones más claras (Loisel et al., 2019).

De esta manera, este estudio confirma lo señalado en la literatura internacional: los modelos predictivos más avanzados suelen alcanzar mejores métricas de pronóstico. Sin embargo, también resalta la necesidad de mantener un enfoque explicativo, especialmente en sectores regulados como

el asegurador, donde comprender las causas es tan relevante como anticipar los resultados. Al mismo tiempo, la incorporación de evidencia latinoamericana otorga un valor añadido, al mostrar que los determinantes de la fuga pueden variar según el contexto local. Esto sugiere que las políticas de retención diseñadas únicamente a partir de estudios internacionales podrían resultar insuficientes si no se adaptan a las particularidades del mercado más local.

7.3. Importancia para la empresa

En cuanto al ámbito empresarial, se tiene que más allá de la contribución académica, los resultados de este estudio tienen una relevancia práctica directa para la gestión de la aseguradora analizada. El hecho de contar con dos modelos complementarios, uno con capacidad explicativa (logit) y otro con alto poder predictivo (red neuronal), abre la posibilidad de diseñar políticas de retención más sofisticadas, basadas tanto en la comprensión de los factores que impulsan la fuga como en la identificación temprana de clientes en riesgo.

Como ya se ha mencionado anteriormente dentro de este documento, desde la perspectiva estratégica, el modelo logit permite a la empresa comprender qué variables influyen significativamente en la probabilidad de abandono. Por ejemplo, la asociación positiva entre morosidad y fuga, así como el efecto estabilizador de la antigüedad contractual, ofrecen insumos concretos para definir políticas de fidelización. La literatura en gestión de clientes ha demostrado que el diseño de programas de retención focalizados en segmentos de alto riesgo puede generar incrementos sustanciales en rentabilidad y reducir el costo de adquisición de nuevos asegurados (Reichheld and Sasser, 1990; Gallo, 2014). Así, los resultados obtenidos no solo validan estas tendencias en el contexto chileno, sino que además las cuantifican en términos específicos para la compañía,

entregando evidencia de cuáles son los segmentos prioritarios a abordar.

Por otra parte, la red neuronal se presenta como una herramienta clave para la operación diaria. Su capacidad de mejorar métricas como el *recall* y el F1-score permite desarrollar sistemas de alerta temprana que identifiquen de manera más precisa a los clientes con mayor probabilidad de fuga. Con esta información, la empresa puede implementar acciones preventivas, tales como ofertas personalizadas, ajustes en la comunicación comercial o beneficios adicionales orientados a reducir la propensión a abandonar. De acuerdo con un reporte sectorial, cerca del 78 % de las aseguradoras a nivel mundial están usando o planean implementar modelos predictivos de fuga en los próximos dos años, y aquellas que han aplicado estas herramientas han logrado incrementos promedio de 23 % en el valor de vida del cliente (InsuredMine, 2024)[6].

En definitiva, los resultados de este estudio ofrecen a la aseguradora una oportunidad concreta de fortalecer su estrategia de fidelización. Al disponer de evidencia sobre los factores que inciden en la fuga y de herramientas predictivas para anticipar casos de riesgo, la empresa puede orientar mejor sus recursos comerciales hacia los segmentos más vulnerables, reducir los costos asociados a la pérdida de clientes y mejorar la efectividad de sus campañas de retención. En un mercado caracterizado por márgenes decrecientes y una alta presión regulatoria, contar con este tipo de capacidades analíticas se convierte en un elemento importante para asegurar la sostenibilidad y competitividad de la compañía en el largo plazo.

7.4. Líneas de investigación futuras

A partir de estos hallazgos, se abre la posibilidad de explorar diversas líneas de investigación futura, especialmente orientadas a la aplicación práctica dentro de la empresa y el sector

[6]Fuente de carácter industrial, no académico, publicada en el blog corporativo de InsuredMine (2024).

asegurador chileno. En primer lugar, se puede apuntar a la incorporación de variables externas que complementen la información administrativa de la compañía. Variables macroeconómicas, como indicadores de desempleo, inflación o ciclos económicos, ya que estos pueden tener un efecto relevante en la decisión de cancelar o renovar pólizas, lo que hace recomendable explorar su integración en futuros modelos predictivos.

En segundo lugar, la aplicación de técnicas de aprendizaje automático puede ampliarse a otros problemas críticos para la gestión de riesgos en la industria aseguradora. Un ejemplo relevante es la detección de fraude en reclamos, donde modelos similares han demostrado mejoras significativas en precisión y ahorro de costos (Ngai et al., 2011). De esta forma, la experiencia adquirida en el modelado de fuga puede ser reutilizada para abordar problemáticas complementarias que también afectan la rentabilidad de la empresa.

Por otro lado, se recomienda explorar la combinación de modelos predictivos con herramientas de interpretabilidad basadas en inteligencia artificial explicable (XAI). El uso de técnicas como SHAP o LIME permitiría aumentar la transparencia de los modelos complejos, equilibrando el desempeño predictivo de algoritmos como redes neuronales con la necesidad de explicaciones claras para la toma de decisiones (Molnar, 2020). De esta manera, se avanzaría hacia modelos que no solo sean más robustos en términos técnicos, sino también más confiables y aceptados dentro de la organización.

Adicionalmente, una línea de trabajo especialmente relevante consiste en la automatización del proceso analítico y de generación de pronósticos. La implementación de un sistema que se alimente periódicamente de la base de datos corporativa, limpiando, estandarizando y actualizando la información de forma automática, permitiría que el modelo se ejecute de manera programada, por ejemplo, cada noche. De este modo, al comenzar la jornada laboral, la empresa podría disponer

de estimaciones actualizadas de probabilidad de fuga y reportes de riesgo listos para su análisis. Este tipo de automatización convertiría el modelo en una herramienta de alerta temprana integrada a los flujos operativos, con capacidad de adaptación continua ante la evolución de los datos reales.

Finalmente, es pertinente abrir una reflexión sobre las implicancias laborales y organizacionales asociadas a la adopción de estos sistemas inteligentes. La automatización de tareas analíticas y la delegación parcial del juicio predictivo a algoritmos generan nuevos desafíos para el rol del ingeniero civil industrial y de los profesionales de la gestión. Ya que en un contexto donde los procesos pueden funcionar de manera autónoma, el valor agregado del ingeniero radica menos en la ejecución rutinaria y más en su capacidad para interpretar modelos, supervisar su funcionamiento y traducir los resultados en decisiones estratégicas. Este fenómeno plantea interrogantes sobre la demanda futura de perfiles profesionales, la redefinición de competencias y los impactos estructurales que la inteligencia artificial puede tener en la vida empresarial. Por lo que, el hecho de profundizar en estos aspectos representa una oportunidad de investigación interdisciplinaria para comprender cómo la automatización va a transformar la práctica profesional y la organización del trabajo en la industria aseguradora y en muchas empresas más de rubros similares.

8. Limitaciones

Tal y como se ha indicado a lo largo del trabajo, la principal limitación metodológica proviene de los modelos utilizados. El modelo logit permite una interpretación directa de los factores que influyen en la fuga, aunque no es capaz de capturar interacciones de carácter no lineal entre las variables. La red neuronal, en contraposición, posee una alta capacidad predictiva, pero su funcionamiento interno es menos transparente, lo que dificulta comprender de manera exacta cómo cada variable contribuye al resultado final. En síntesis, ambos modelos ofrecen un equilibrio razonable entre capacidad predictiva e interpretabilidad, suficiente para cumplir con los objetivos de la investigación, aunque sus resultados siguen siendo una aproximación a una realidad que siempre puede perfeccionarse con nuevas fuentes de información y metodologías más avanzadas.

Otra restricción importante a considerar es el ámbito de aplicación. Este estudio se centra en los registros históricos de una única aseguradora, lo que garantiza que los hallazgos sean relevantes para la empresa en cuestión, pero limita la capacidad de aplicar estos resultados a toda la industria. En otros contextos, las dinámicas competitivas, las regulaciones o incluso las diferencias culturales pueden influir en los factores que llevan a la fuga de clientes.

Además, la base de datos utilizada se compone principalmente de información contractual y sociodemográfica, dejando de lado variables relacionadas con el comportamiento digital o factores macroeconómicos. Esta limitación dificulta la exploración de fenómenos más amplios, como el impacto del ciclo económico, la competencia entre aseguradoras o cómo las campañas de marketing digital afectan la retención de clientes.

En cuanto a la medición del desempeño, aunque se utilizaron métricas adecuadas para conjuntos de datos desbalanceados, como el recall y el F1-score, no se realizó una evaluación del

impacto económico directo de las predicciones. En la realidad, las compañías de seguros no solo necesitan identificar con precisión a los clientes en riesgo, sino también calcular el valor financiero que implica retenerlos. Este aspecto podría ser explorado en estudios futuros a través de modelos de costo-beneficio o simulaciones de retorno esperado.

Además, aunque la combinación de logit y redes neuronales se alineó con los objetivos del estudio, la adición de otros algoritmos, como Random Forest, XGBoost o técnicas de aprendizaje por ensamblado, podría potenciar la capacidad predictiva o brindar perspectivas adicionales. Sin embargo, dentro del marco establecido, las metodologías empleadas ofrecen resultados sólidos y valiosos para la gestión de la aseguradora.

Finalmente, los modelos se entrenaron utilizando información histórica. Esto significa que futuros cambios, como la llegada de aseguradoras digitales, ajustes regulatorios o cambios en el comportamiento del consumidor, podrían alterar las relaciones observadas y hacer que los resultados pierdan relevancia. Por lo tanto, es importante interpretar las conclusiones dentro del contexto temporal y organizacional analizado, manteniendo la mente abierta a actualizaciones a medida que el entorno competitivo y tecnológico evoluciona.

Además, es relevante señalar que la implementación de modelos de inteligencia artificial en el sector asegurador presenta desafíos relacionados con la transparencia y la ética en el manejo de datos. Avanzar hacia modelos más comprensibles y auditables es un paso crucial para fortalecer la confianza y garantizar un uso responsable de la analítica predictiva.

En conjunto, estas limitaciones no invalidan los resultados obtenidos, pero sí orientan futuras mejoras metodológicas y de disponibilidad de información. Reconocerlas permite situar adecuadamente el alcance de las conclusiones y destacar que los hallazgos presentados constituyen una buena base sobre la cual profundizar en investigaciones posteriores que incorporen nuevas

variables, metodologías o contextos del mercado asegurador chileno.

9. Conclusiones

El objetivo de este trabajo fue modelar y predecir la fuga de clientes en el sector asegurador chileno, comparando métodos econométricos tradicionales (logit y probit) con técnicas de aprendizaje automático, en particular redes neuronales multicapa. A partir de una base de datos anonimizada entregada por la aseguradora, se buscó identificar los factores que influyen en la cancelación de pólizas y evaluar qué tan bien distintos modelos permiten anticipar este comportamiento.

Los resultados mostraron que los modelos logit, aunque tienen un menor desempeño al momento de predecir, entregan información muy valiosa para entender las causas detrás de la fuga. Variables como la antigüedad del contrato, la existencia de documentos impagos y el monto de la prima básica resultaron ser factores relevantes que aumentan o reducen la probabilidad de cancelación. Este tipo de información permite respaldar decisiones estratégicas con evidencia concreta. Por otro lado, las redes neuronales demostraron una capacidad mucho mayor para anticipar qué clientes están en riesgo de fuga, alcanzando buenos resultados en métricas como *recall*, F1 y AUC. Esto las convierte en una herramienta útil para diseñar sistemas de alerta temprana y programas de retención más precisos.

Por lo anterior, ambos enfoques no deben verse como opuestos, sino como complementarios. El modelo logit ayuda a comprender los factores que influyen en la decisión de fuga, mientras que la red neuronal permite anticipar con más acierto los posibles casos. En conjunto, entregan una visión completa que combina explicación y predicción, algo clave en una industria tan competitiva y regulada como la de seguros.

Desde el punto de vista académico, esta memoria aporta evidencia para la región latinoamericana, donde aún existen pocos estudios sobre predicción de fuga en el sector asegurador. Se

demuestra que combinar modelos explicativos con modelos predictivos enriquece tanto el análisis de los datos como su aplicación práctica. En el plano empresarial, los resultados entregan a la aseguradora una herramienta concreta que puede replicarse para mejorar la retención de clientes, optimizar recursos comerciales y aumentar la efectividad de las campañas, alineándose con las tendencias actuales de gestión basada en datos.

El uso conjunto de modelos econométricos y de aprendizaje profundo no solo mejora la capacidad analítica de la empresa, sino que también abre nuevas oportunidades de aplicación. Entre ellas, la detección de fraude, la evaluación de riesgos y la incorporación de variables externas, como indicadores macroeconómicos o de comportamiento, que podrían fortalecer aún más las predicciones futuras.

Finalmente, este trabajo refuerza el rol del ingeniero civil industrial como un profesional capaz de integrar análisis cuantitativo, herramientas tecnológicas y una visión estratégica para resolver problemas reales dentro de las organizaciones. La incorporación responsable de la inteligencia artificial en el ámbito asegurador no solo representa una oportunidad de innovación, sino también un compromiso ético con la toma de decisiones informadas y transparentes. En conjunto, los resultados obtenidos demuestran que la analítica predictiva, aplicada con criterio y responsabilidad, puede transformarse en un pilar clave para la competitividad y el desarrollo sostenible de la industria de seguros en Chile.

10. Anexos

A. Notas metodológicas complementarias

Esta sección reúne aspectos técnicos de la estimación y validación de los modelos econométricos utilizados en la investigación. Se incluyen procedimientos de máxima verosimilitud, pruebas de significancia, controles de identificabilidad y verificaciones de robustez numérica.

Conceptos utilizados

Estimación por máxima verosimilitud y Newton-Raphson. El modelo logit se estima maximizando la log-verosimilitud $\ell(\beta) = \sum_i \left[y_i \log p_i + (1 - y_i) \log(1 - p_i)\right]$, con $p_i = \sigma(x_i'\beta)$. La maximización se realiza mediante Newton–Raphson (IRLS), que actualiza iterativamente los coeficientes utilizando el gradiente y la curvatura de $\ell(\beta)$ hasta la convergencia. Este procedimiento es estándar y eficiente en presencia de múltiples variables indicadoras.

Efectos marginales promedio (AME). Además de los coeficientes (en escala de log-odds), se informan los AME, que cuantifican el cambio promedio en la probabilidad de fuga ante una variación unitaria del regresor (o de $0 \to 1$ si es dicotómica). Facilitan la interpretación económica de los resultados.

Test de razón de verosimilitudes (LR). Para contrastar modelos anidados estimados sobre la misma muestra, se utiliza la estadística:

$$\mathrm{LR} = 2\big(\ell_{\text{grande}} - \ell_{\text{chico}}\big) \;\sim\; \chi^2_{\text{df}},$$

donde *df* es la diferencia en el número de parámetros entre ambos modelos. Es decir, de M1 a M2 se agrega `EDAD`, `EDAD`^{2} y `SEXO` (*df*=3); luego de M2 a M3 se agrega `DOC_IMPAGOS`, `ANTIGUEDAD_MESES`, `BENEFICIOS_RUT` y `Nueva_MNPMP` (*df*=4). Valores altos de LR con $p < 0{,}001$ en los modelos evaluados indican que el añadir variables mejoró significativamente el ajuste.

Test de Wald (contrastes por bloques). Dentro de un modelo dado, se evalúan hipótesis conjuntas del tipo $H_0 : \beta_{j_1} = \cdots = \beta_{j_q} = 0$ mediante el estadístico de Wald, $W \sim \chi^2_q$. Este contraste se utiliza para verificar el aporte global de grupos de variables (por ejemplo, características individuales en M2 o variables de relación con la aseguradora en M3).

Controles de identificabilidad y estabilidad numérica. Se excluyen automáticamente variables dicotómicas sin variación o con predicción perfecta, que impiden la estimación (coeficientes no acotados). El centrado de `EDAD` y la inclusión del término cuadrático mejoran la estabilidad y permiten capturar no linealidades.

Reestimaciones de robustez y "jitter". Para descartar dependencia de los valores iniciales, se reestiman las especificaciones partiendo (i) desde ceros y (ii) con pequeñas perturbaciones numéricas (jitter) en los coeficientes de las variables recién añadidas al pasar de un modelo al siguiente. Las log-verosimilitudes obtenidas son virtualmente idénticas a las de la estimación principal, lo que respalda la robustez de los resultados.

Procedimiento de estimación y validaciones. Los modelos se estiman por máxima verosimilitud (logit) usando el método de Newton; entre especificaciones se utiliza warm-start únicamente

para acelerar la convergencia (sin afectar el óptimo). Para facilitar la interpretación se reportan, junto a los coeficientes, los efectos marginales promedio (AME). La mejora al ampliar el conjunto de regresores se contrasta con tests de razón de verosimilitud y con tests de Wald por bloques. También se verifica el rango de las matrices de diseño y se realizan reestimaciones de control (sin warm-start y con jitter en las variables nuevas), sin cambios sustantivos.

B. Código de limpieza de datos

A continuación, se presenta el script en `Python` utilizado para la depuración, transformación y generación de la base de datos final utilizada.

Listing 1: Script de limpieza y preparación de la base de datos

```
# LIBRERÍAS UTILIZADAS
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pingouin as pg
import statsmodels.api as sm
import itertools
from scipy import stats
from itertools import combinations
from matplotlib.gridspec import GridSpec
from scipy.stats import randint, pearsonr, chi2_contingency, ks_2samp,
    pointbiserialr, jarque_bera, kstest, shapiro, t, chi2, chisquare, poisson,
     kstest, norm
from sklearn.model_selection import train_test_split, GridSearchCV,
    RandomizedSearchCV
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression, LinearRegression
from sklearn.preprocessing import LabelEncoder, StandardScaler, OneHotEncoder
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
```

```
from sklearn.metrics import accuracy_score, confusion_matrix,
    classification_report, precision_recall_curve, precision_score,
    recall_score, f1_score, r2_score
from sklearn.metrics import mean_squared_error
from datetime import datetime, timedelta
from statsmodels.stats.outliers_influence import variance_inflation_factor
from sklearn.decomposition import PCA
from sklearn.preprocessing import PolynomialFeatures, StandardScaler,
    OneHotEncoder
from sklearn.pipeline import Pipeline, make_pipeline
from statsmodels.stats.diagnostic import het_breuschpagan
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.ensemble import RandomForestClassifier,
    GradientBoostingClassifier
from sklearn.svm import SVC
from joblib import dump, load
plt.rcParams['figure.figsize']=(10, 6)
import warnings
warnings.filterwarnings('ignore')
# LEER LA BASE PRIMERO
base='Base modelo 2 nuevos datos.xlsx'
basedatos=pd.read_excel(base)

# Función para asignar valores dicotómicos a la columna 'STPOL'
def dicotomica_estado(categoria):
    if "A" in categoria:
        return 1
```

```
    elif "V" in categoria:
        return 0
    else:
        return None

# Función para asignar valores dicotómicos a la columna 'STSEXO'
def dicotomica_edad(categoria):
    if pd.notna(categoria): # Verificar si el valor no es NaN
        if "F" in categoria:
            return 1
        elif "M" in categoria:
            return 0
    return None # para valores NaN

# LIMPIEZA DE BASE DE DATOS, SE ELIMINAN REGISTROS CON VALORES FALTANTES:
basedatos = basedatos[basedatos['STPOL'] != 'D'].reset_index(drop=True) #
    elimina filas con estado poliza "D" y además restablece índices.
basedatos = basedatos[basedatos['MNPMP'] != 0].reset_index(drop=True) #
    elimina filas con monto prima 0 y además restablece índices.
condiciones_eliminar = (basedatos['FCANULA'] == 0) & (basedatos['STPOL'] == 'A
    ') # si está anulada, entonces debe tener fecha anulación, si no se cumple
     esto, se elimina.
basedatos = basedatos.loc[~condiciones_eliminar]
basedatos.reset_index(drop=True, inplace=True)
basedatos = basedatos.dropna(subset=['MNCAAS','MNPMP','MNPRBA','FCNACI','
    FCANULA','DSTCON','SIGLA']) # elimina los registros con valores NaN.
basedatos = basedatos[basedatos['NRPLAN'] != 0].reset_index(drop=True) #
```

```
    elimina filas con PLAN igual a 0 y además restablece índices.
65
66 basedatos['DOC_IMPAGOS'] = basedatos['DOC_IMPAGOS'].fillna(0) # rellena con 0'
    s las celdas vacías.
67 basedatos['BENEFICIOS_RUT'] = basedatos['BENEFICIOS_RUT'].fillna(0) # rellena
    con 0's las celdas vacías.
68
69 # Acá se aplican las condiciones y se crea una nueva columna 'Nueva_MNPMP':
70 basedatos['Nueva_MNPMP'] = basedatos.apply(lambda row: 0 if row['MNPMP'] ==
    row['MNPRBA'] else row['MNPRBA'] - row['MNPMP'], axis=1)
71 basedatos = basedatos[basedatos['Nueva_MNPMP'] >= 0].reset_index(drop=True)
72 print(basedatos.head(100)) # imprimen los primeros 100 registros.
73
74 # CÓDIGO QUE CREA LAS VARIABLES y="FUGA" Y "SEXO" A UTILIZAR EN EL MODELO
75 basedatos['STMOAN2'] = basedatos['STMOAN2'].str.replace(r'[\s\t]+', '', regex=
    True)
76 basedatos['STMOAN2'] = basedatos['STMOAN2'].fillna(0).astype(str)
77
78 # Crear una nueva columna 'FUGA' que contiene las categorías consolidadas
79 basedatos['FUGA'] = basedatos['STMOAN2'].replace({
80     'FALLECIMIENTO': 0,
81     'NOEXISTE': 0,
82     'NOPAGO': 1,
83     'RESCATE': 1,
84     'VENCIMIENTO': 0,
85     'RENUNCIA': 1,
86     'FALSIFICACION': 0,
```

```
    'BENEFCESANTIA': 0,
    '0': 0
})

# Aplicar función a la columna 'STSEXO' y crear una nueva columna 'SEXO' con valores 0 y 1:
basedatos['SEXO'] = basedatos['STSEXO'].apply(dicotomica_edad)
basedatos = basedatos.dropna(subset=['SEXO']) # elimina los registros con valores NaN en columna "SEXO".
# OBTENER EDAD: Convertir la columna 'FCNACI' a formato datetime y manejar fechas inválidas
basedatos['FCNACI'] = pd.to_datetime(basedatos['FCNACI'], format='%Y%m%d', errors='coerce')
basedatos['FCEMISIO'] = pd.to_datetime(basedatos['FCEMISIO'], format='%Y%m%d', errors='coerce')
edad_dias = (basedatos['FCEMISIO'] - basedatos['FCNACI']).dt.days # Calcula la diferencia en días entre la fecha actual y la fecha de nacimiento convertida.

# Crear una nueva columna 'EDAD' con valor predeterminado (0):
basedatos['EDAD'] = 0
# Aplicar lógica condicional para calcular la edad solo cuando la fecha es válida (CON NOTNULL):
condicional = basedatos['FCNACI'].notnull()
# Si la columna 'FCNACI' no es nula (notnull()), entonces se calcula la edad en años dividiendo la diferencia en días por 365.25 para tener en cuenta los años bisiestos:
```

```
basedatos.loc[condicional, 'EDAD'] = ((basedatos['FCEMISIO'] - basedatos['
    FCNACI']).dt.days / 365.25).round()
basedatos['EDAD'] = basedatos['EDAD'].astype(int) # redondea las edades y las
    convierte a int.
basedatos = basedatos[basedatos['EDAD'] >= 18].reset_index(drop=True) #
    eliminar filas edades menor a 18 años.

#----------------------------#
# OBTENER ANTIGUEDAD DE ASEGURADOS (MESES)
#----------------------------#
# Convertir las columnas 'FCINCON' y 'FCANULA' a formato datetime:
basedatos['FCEMISIO'] = pd.to_datetime(basedatos['FCEMISIO'], format='%Y%m%d'
    , errors='coerce')
basedatos['FCANULA'] = pd.to_datetime(basedatos['FCANULA'], format='%Y%m%d',
    errors='coerce')
fecha_actual = datetime.now() # fecha actual
basedatos['FCANULA'].fillna(fecha_actual, inplace=True) # Reemplazar celdas
    NaN en 'FCANULA' con la fecha actual

# Calcular la antigüedad en días:
basedatos['ANTIGUEDAD_DIAS'] = (basedatos['FCANULA'] - basedatos['FCEMISIO']).
    dt.days
basedatos['ANTIGUEDAD'] = (basedatos['ANTIGUEDAD_DIAS']/30)

# Calcular la antigüedad en meses:
basedatos['ANTIGUEDAD_MESES'] = (fecha_actual.year - basedatos['FCEMISIO'].dt.
    year) * 12 + fecha_actual.month - basedatos['FCEMISIO'].dt.month
```

```
123
124 # Calcular la antigüedad en meses para pólizas anuladas:
125 basedatos.loc[basedatos['FCANULA'].notnull(), 'ANTIGUEDAD_MESES'] = ((
        basedatos['FCANULA'] - basedatos['FCEMISIO']).dt.days / 30).astype(int)
126
127 # ELIMINAR ANTIGUEDAD DEL CLIENTE (MESES) QUE SEAN NEGATIVOS, ESTÁN MAL
        REGISTRADOS LOS DATOS
128 basedatos = basedatos[basedatos['ANTIGUEDAD_DIAS'] >= 0] # filtra filas con
        ANTIGUEDAD_DIAS negativos y se eliminan, dejamos los positivos solamente.
129 basedatos.reset_index(drop=True, inplace=True) # restablece índices después de
         eliminar filas.
130
131 #-----------------------------#
132 # Creación de XX variables dicotómicas para señalar los XX productos (seguros)
133 #-----------------------------#
134 # Crear variables dicotómicas para cada seguro
135 seguros = pd.get_dummies(basedatos['NRPLAN'], prefix='SEGURO', dtype=float,
        drop_first=True)
136 # Concatenar las nuevas columnas al DataFrame original
137 basedatos = pd.concat([basedatos, seguros], axis=1)
138
139 #-----------------------------#
140 # Creación de XX variables dicotómicas para señalar los XX canales de venta
141 #-----------------------------#
142 # Crear variables dicotómicas para cada intermediario
143 intermediarios = pd.get_dummies(basedatos['DSTCON'], prefix='INTERMEDIARIO',
        dtype=float, drop_first=True)
```

```
144 # Concatenar las nuevas columnas al DataFrame original
145 basedatos = pd.concat([basedatos, intermediarios], axis=1)
146
147 #-----------------------------#
148 # Creación de XX variables dicotómicas por zona en donde se vendió el seguro
149 #-----------------------------#
150 # Crear variables dicotómicas para cada zona
151 zonas = pd.get_dummies(basedatos['SIGLA'], prefix='ZONA', dtype=float,
        drop_first=True)
152 # Concatenar las nuevas columnas al DataFrame original
153 basedatos = pd.concat([basedatos, zonas], axis=1)
154
155 #-----------------------------#
156 # BASE OBTENIDA
157 #-----------------------------#
158 # Si es necesario analizar alguna poliza en específico, añadir este campo: '
        NRPOLI' al comienzo de la siguiente línea de código
159 columnas_utilizadas = ['NRPOLI','FUGA','EDAD','SEXO','ANTIGUEDAD_MESES','
        MNPRBA','Nueva_MNPMP','MNCAAS','DOC_IMPAGOS','BENEFICIOS_RUT']
160 # Agregar las variables dicotómicas de productos:
161 columnas_utilizadas.extend(seguros.columns)
162 # Agregar las variables dicotómicas de intermediarios:
163 columnas_utilizadas.extend(intermediarios.columns)
164 # Agregar las variables dicotómicas de zonas:
165 columnas_utilizadas.extend(zonas.columns)
166 # Selección de las columnas necesarias para el modelo:
167 basedatos_resultado = basedatos[columnas_utilizadas]
```

```
168 # Guardar el DataFrame resultante en un archivo Excel:
169 basedatos_resultado.to_excel('BASE_FINAL.xlsx', index=False)
```

Referencias

AbdelAziz, N. M., Bekheet, M., Salah, A., El-Saber, N., and AbdelMoneim, W. T. (2025). A comprehensive evaluation of machine learning and deep learning models for churn prediction. *Information*, 16(7):537.

Ahn, J.-H., Park, S.-Y., and Kim, D. (2024). Hybrid black-box classification for customer churn: Combining interpretability and performance. *Decision Support Systems*, 177:113966.

AM Best (2025). Am best maintains stable outlook on chile's insurance industry. BestWire, 8 de mayo de 2025.

Asociación de Aseguradores de Chile (AACh) (2024). Síntesis estadística del mercado asegurador 2024. (Consultado el 17 de agosto de 2025).

Azzone, M., Barucci, E., Guerrero Gómez, G., and Marazzina, D. (2022). A machine learning model for lapse prediction in life insurance contracts. *Expert Systems with Applications*, 191:116261.

Bain & Company (2014). The economics of e-loyalty. Executive Brief / Idea in Brief. (Consultado el 20 de agosto de 2025).

Bogaert, M. and Delaere, L. (2023). Ensemble methods in customer churn prediction: A comparative analysis of the state-of-the-art. *Mathematics*, 11(5):1137.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636.

Castro, J. P., Herrera, C., and Ramírez, E. (2023). Customer churn prediction in emerging markets: Evidence from latin america. *Journal of Business Research*, 161:113849.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 321–327. ACM.

Comisión para el Mercado Financiero (CMF) (2024). Informe financiero del mercado asegurador a septiembre 2024. (Consultado el 17 de agosto de 2025).

Coussement, K. and Van den Poel, D. (2008). Integrating the voice of customers through call center emails into a decision support system for churn prediction. *Information & Management*, 45(3):164–174.

CustomerGauge (2024). Average churn rate by industry (2024 update). Consultado el 21 de agosto de 2025.

Dong, Y., Frees, E. W., Huang, F., and Hui, F. K. C. (2022). Multi-state modelling of customer churn. *ASTIN Bulletin: The Journal of the IAA*, 52(3):735–764.

D'Espallier, B., Guérin, I., and Mersland, R. (2011). Women and repayment in microfinance: A global analysis. *World Development*, 39(5):758–772.

European Insurance and Occupational Pensions Authority (EIOPA) (2024). Report on the digitalisation of the european insurance sector. Technical report, EIOPA. Accessed 3 Nov 2025.

Gallo, A. (2014). The value of keeping the right customers. *Harvard Business Review*. (Consultado el 20 de agosto de 2025).

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. Consultado el 21 de agosto de 2025.

Gupta, S., Lehmann, D. R., and Stuart, J. A. (2006). Customer lifetime value research in marketing: A review and future directions. *Journal of Interactive Marketing*, 20(2):61–74.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1322–1328.

Henao Madrigal, M., Restrepo Tobón, D., and Laniado, H. (2020). Customer churn prediction in insurance industries: A multiproduct approach. *Working Paper*.

InsuredMine (2024). From risk to retention: Transforming insurance with predictive churn analytics. (Consultado el 21 de agosto de 2025).

Lemmens, A. and Croux, C. (2006). Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, 43(2):276–286.

Liu, X., Xia, G., Zhang, X., Ma, W., and Yu, C. (2024). Customer churn prediction model based on hybrid neural networks (ccp-net). *Scientific Reports*, 14(30707).

Loisel, S., Piette, P., and Tsai, J. (2019). Applying economic measures to lapse risk management with machine learning approaches. *arXiv preprint arXiv:1906.05087*.

Lorca Figueroa, S. P. (2021). Predicción de fuga de clientes mediante el rediseño del proceso de atención en seguros masivos para una compañía aseguradora. Master's thesis, Universidad de Chile.

Manteigas, C. and António, N. (2024). Understanding and predicting lapses in mortgage life insurance using a machine learning approach. *Expert Systems with Applications*, 255:124753.

Mena, C. G., De Caigny, A., Coussement, K., De Bock, K. W., and Lessmann, S. (2019). Churn prediction with sequential data and deep neural networks: A comparative analysis. *arXiv preprint arXiv:1909.11114*.

Mengistu, A., Tesfaye, M., and Kebede, F. (2022). Customer churn prediction using machine learning techniques: The case of lion insurance. Preprint. Versión ampliamente difundida en Zenodo/ResearchGate; no se encontró edición IJCA con volumen/número verificables.

Molnar, C. (2020). *Interpretable Machine Learning*. Leanpub, 2nd edition. Consultado el 21 de agosto de 2025.

Ngai, E. W., Hu, Y., Wong, Y., Chen, Y., and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569.

Organisation for Economic Co-operation and Development (OECD) (2023). Leveraging technology in insurance to enhance risk assessment and policyholder risk reduction. Technical report, OECD. Accessed 3 Nov 2025.

Prabadevi, B., Shalini, R., and Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks*, 4(May):145–154.

Reck, L., Schupp, J., and Reuß, A. (2023). Identifying the determinants of lapse rates in life insurance: an automated lasso approach. *European Actuarial Journal*, 13(1):149–178.

Reichheld, F. F. and Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review*, 68(5):105–111. Consultado el 21 de agosto de 2025.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85.

Sharma, A. (2013). A neural network based approach for predicting customer churn. *International Journal of Computer Applications*, 27(11):42–47.

Society of Actuaries (2021). Interpretable machine learning for insurance applications. Technical report, Society of Actuaries. Consultado el 21 de agosto de 2025.

StataCorp LLC (2025). *logit — Logistic regression, reporting coefficients*. StataCorp LLC. Methods and formulas; likelihood and references.

Valla, M., Chambaz, A., Curis, E., and Donnat, C. (2024). A longitudinal tree-based framework for lapse management in life insurance. *Analytics*, 3(3):318–343.

Verbeke, W., Martens, D., Mues, C., and Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3):2354–2364.

Wooldridge, J. M. (2009). *Introducción a la Econometría: Un Enfoque Moderno*. Cengage Learning, México, 4a edition.

Yin, S., Dey, D. K., Valdez, E. A., Gan, G., and Vadiveloo, J. (2020). Skewed link regression models for imbalanced binary response with applications to life insurance. arXiv preprint.