Universidad Internacional de La Rioja

Facultad de Empresa, Comunicación y Marketing

Máster Universitario en Inteligencia de Negocio

Predicción del churn de influencers en stylink: Modelos de clasificación para optimizar la retención

Trabajo fin de estudio presentado por:	Katia Mogán Roncero
Tipo de trabajo:	Proyecto de Inteligencia de Negocio
Modalidad:	Individual
Director/a:	Juan David Bohórquez Giraldo
Fecha:	04.07.2025

Resumen

El éxito de una empresa depende en gran medida de su capacidad para adquirir nuevos clientes mientras mantiene a los existentes. Las organizaciones con altas tasas de retención se benefician de un crecimiento constante, lo que también puede generar un mayor número de referencias por parte de sus clientes actuales.

El churn ocurre cuando un cliente deja de tener relación con la empresa. Predecir la pérdida de estos es importante para las organizaciones, ya que retener a los clientes existentes suele ser menos difícil y costoso que adquirir nuevos.

En este estudio se plantea un modelo de predicción de churn enmarcado en el contexto de la industria del marketing de afiliación. Los datos recopilados proceden de stylink, una empresa alemana especializada en marketing de afiliación con y para influencers. Estos incluyen información sobre los propios influencers, enlaces de productos creados y datos de compra.

Con el objetivo de estimar la probabilidad de churn, se emplean dos técnicas de aprendizaje supervisado: la regresión logística y los árboles de decisión. La bondad de los modelos se evalúa mediante diferentes métricas y, con base en los resultados, se proponen estrategias de retención para contrarrestar los efectos del aumento de competencia en este mercado.

Palabras clave: predicción churn de cliente, marketing de influencers, marketing de afiliación, estrategia de cliente

Abstract

A company's success largely depends on its ability to acquire new customers while retaining existing ones. Organizations with high retention rates benefit from consistent growth, which can also lead to a higher number of referrals from their current customers.

Churn occurs when a customer ceases their relationship with a company. Predicting customer churn is important for organizations, as retaining existing customers is often less difficult and costly than acquiring new ones.

This study presents a churn prediction model within the context of the affiliate marketing industry. The data collected comes from stylink, a German company that specializes in end-to-end affiliate marketing with and for influencers. This data includes information about influencers, created product links, and purchasing behavior.

To estimate the probability of churn, two supervised machine learning techniques are implemented: logistic regression and decision trees. The goodness of the models is evaluated using different metrics, and based on the results, retention strategies are proposed to counteract the effects of increased competition in this market.

Keywords: customer churn prediction, influencer marketing, affiliate marketing, customer-based strategy

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

Dado que retener a un cliente es seis veces más barato que adquirir uno nuevo, el éxito de una empresa depende en gran medida de su capacidad para atraer nuevos clientes mientras retiene a los existentes. Las industrias que se benefician de relaciones a largo plazo con los clientes—como el comercio electrónico, los servicios financieros y las aplicaciones móviles—buscan activamente formas de retener a sus clientes e identificar los factores que conducen al churn (Kim & Lee, 2021).

El churn se refiere a la pérdida de clientes que dejan de utilizar un servicio de la empresa. Puede darse en distintos contextos, como clientes que dejan de realizar compras o suscriptores que cancelan un servicio móvil (Misirlis & Vlachopoulou, 2021). Las causas del churn pueden incluir una baja satisfacción del cliente, el aumento de la competencia, la introducción de nuevos productos y/o marcas, cambios en la normativa y otras dinámicas del mercado.

El presente trabajo se lleva a cabo siguiendo una metodología CRISP-DM (Shearer, 2000) y en colaboración con stylink, una empresa alemana fundada en 2019, con sede central en Münster y presencia en 13 países. stylink se especializa en el marketing de afiliación y UGC (User Generated Content), y actúa como intermediaria entre influencers y marcas de diversos sectores, como el de la moda, decoración o electrónica.

Este estudio se enmarca en el contexto del comercio electrónico, en un nicho de mercado específico, ya que stylink ofrece servicios tanto en un modelo B2B (stylink – redes de afiliación – marcas) como en un modelo B2C (stylink - influencers), lo que conllevará la necesidad de definir ciertos conceptos clave durante la fase de comprensión del negocio. Si un influencer crea enlaces de productos de manera regular, pero deja de hacerlo durante varios meses, puede considerarse que ha abandonado la plataforma.

El análisis predictivo de churn permitirá a la empresa identificar señales tempranas de abandono y reconocer a los influencers con una mayor probabilidad de dejar de crear enlaces. Este se realiza empleando datos históricos proporcionados por la empresa, y extraídos de sus bases de datos, que contienen información sensible relacionada con su rendimiento, información personal de los influencers, nombres de redes de afiliación y marcas. Por este motivo, la mayoría de los datos se presentarán de forma anonimizada.

1.1. Planteamiento general: descripción y justificación del proyecto

El auge del comercio electrónico, impulsado por la pandemia de COVID-19, ha supuesto un fenómeno sin precedentes en los últimos años, lo que ha transformado el mundo del marketing de influencers y las redes sociales (Kemp, 2021).

En el mercado alemán, stylink ha sido uno de los beneficiarios de este auge, ya que la empresa ha experimentado un crecimiento exponencial en sus ingresos y número de usuarios desde el año 2020.

Sin embargo, los competidores no tardaron en hacerse eco de la popularidad de este tipo de modelo de negocio, lo que, especialmente a partir de 2022, supuso que la empresa enfrentara una creciente pérdida de cuota de mercado y de usuarios quienes, por diversos motivos, empezaron a abandonar stylink en favor de otras plataformas de monetización de contenido que ofrecían servicios similares. No obstante, es difícil definir quiénes son los usuarios que han abandonado la plataforma.

En aras de resolver esta problemática, se utilizarán los datos recopilados por stylink, que incluyen los enlaces creados por los influencers, así como datos transaccionales desde enero de 2024 hasta la fecha. Para predecir el churn de los influencers, se emplearán modelos de clasificación basados en técnicas de aprendizaje supervisado.

Primero, se organizarán y preprocesarán los datos para construir modelos de churn, extrayendo las características más relevantes. Posteriormente, se aplicarán los modelos de regresión logística y árboles de decisión para predecir la probabilidad de que un influencer abandone la plataforma. Finalmente, se realizarán una serie de recomendaciones para mejorar la estrategia de retención de cliente de stylink.

1.2. Objetivos del TFE

1.2.1. Objetivo general

El objetivo principal de este trabajo consiste en analizar en profundidad y predecir la tasa de abandono de los usuarios de stylink. Para ello, se emplearán técnicas de aprendizaje automático capaces de clasificar a estos usuarios según su probabilidad de abandono.

1.2.2. Objetivos específicos

- Realizar un análisis descriptivo para identificar las variables que influyen en el abandono de los usuarios.
- Definir el concepto de "churn" o abandono aplicado al modelo de negocio de stylink, estableciendo las condiciones necesarias para determinar si un influencer ha abandonado la plataforma.
- Extraer, preparar y limpiar los datos para su posterior modelado.
- Utilizar técnicas de aprendizaje supervisado, en concreto la regresión logística y los árboles de decisión, para construir los modelos predictivos. Los datos se dividirán en dos subgrupos: entrenamiento y prueba, con el fin de evaluar la bondad de los modelos mediante diversas métricas de validación.
- Seleccionar el modelo más preciso y aplicarlo a los usuarios para predecir aquellos con mayor riesgo de abandono.
- Ofrecer a stylink recomendaciones para implementar una estrategia de retención adaptada a cada tipo de usuarios según su riesgo de abandono.

1.3. Elementos innovadores del proyecto

La regresión logística y los árboles de decisión son dos técnicas de aprendizaje supervisado comúnmente utilizadas para predecir la tasa de abandono de clientes en diversos sectores (Hassouna, Tarhini, Elyas, & Abou Trab, 2017).

Sin embargo, tras revisar la literatura existente, se observa que los modelos de predicción de churn aplicados al marketing de afiliación son relativamente innovadores.

Este enfoque permitirá a stylink implementar una estrategia de retención y fidelización de clientes más efectiva, lo que incrementará sus ingresos y consolidará su posición en el mercado.

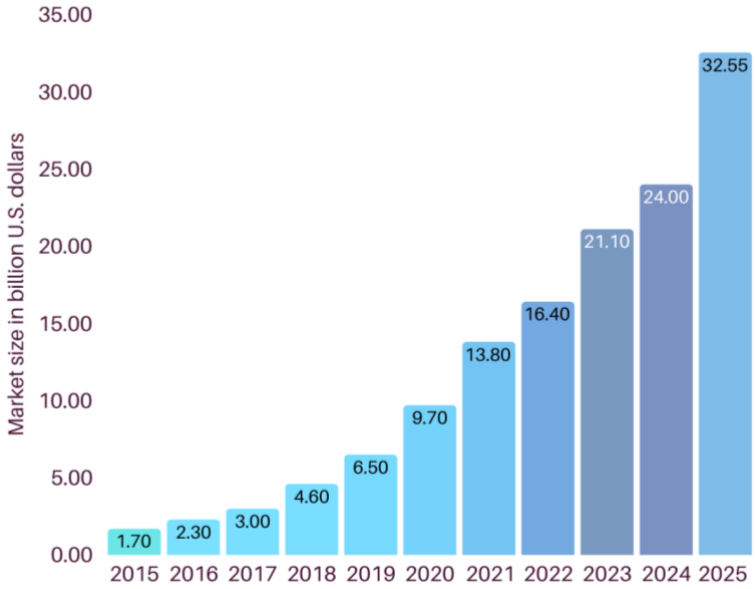
2. Alcance y planificación

2.1. Fase de descubrimiento: evaluación del entorno actual

2.1.1. Marketing de influencers y marketing de afiliación

En los últimos años, el rápido crecimiento de redes sociales como Instagram y TikTok ha transformado la forma en que las empresas se relacionan con los consumidores. Dentro de este panorama digital en evolución, el marketing de influencers y el marketing de afiliación han surgido como estrategias altamente efectivas que las empresas utilizan para contrarrestar la creciente competencia, promocionar sus productos, fidelizar a sus clientes y atraer a nuevas audiencias. Esta tendencia se refleja claramente en la Figura 1, que ilustra el crecimiento exponencial del marketing de influencers en la última década.

Figura 1. *Cuota de mercado mundial del marketing de influencers entre los años 2015 y 2025 (en miles de millones de $ estadounidenses).*



Fuente: Datos de Statista (2024), gráfico recreado por el autor.

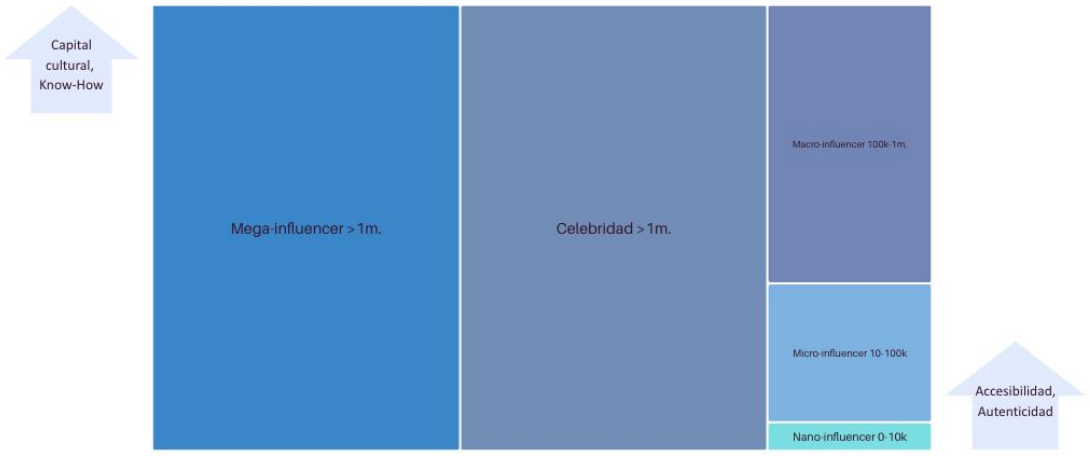
El marketing de influencers es un área del marketing que se enfoca en la promoción de productos o servicios a través de personas influyentes en las redes sociales. Un influencer o 'micro-celebridad' es una persona reconocida que ha obtenido popularidad en redes sociales a través de sus publicaciones, que atraen la atención de miles, cuando no millones de seguidores. Su contenido puede ser de diversa índole: belleza, moda, gastronomía, finanzas, deporte, etc. (Schouten, Janssen, & Verspaget, 2019). La clave de su éxito radica en su

capacidad para interactuar con el público y fidelizar a sus comunidades de seguidores, exactamente lo que las empresas persiguen con sus estrategias de marketing digital. Además, diversos estudios han demostrado que el número de seguidores de los influencers también afecta al rendimiento de estas.

Campbell y Farrell (2020) dividen a los influencers en categorías según su número de seguidores, know-how, capital cultural, accesibilidad y autenticidad percibidas por sus audiencias. Los autores argumentan que, cuanto más pequeño es el influencer, mayor es su accesibilidad y credibilidad y, cuanto más grande, mayor su capital cultural y el know-how que sus audiencias les atribuyen. Existe una literatura amplia con respecto a este tema, pero la gran mayoría coincide en la siguiente clasificación:

- Nano-influencers: 0-10k seguidores. "Newcomers" o influencers noveles.
- Micro-influencers: 10-100k seguidores. "The Rising Star" o estrellas incipientes.
- Macro-influencers: 100k-1 millón seguidores. "The Sweet Spot" o el punto ideal.
- Mega-influencers: +1 millón de seguidores. "The Everyday Celebrity" o el influencer celebrity.
- Celebridad: +1 millón de seguidores. "The Rich and Famous" o la élite, los influencers ricos y famosos (Campbell & Farrell, 2020).

Figura 2. *Tipos de influencer según su número de seguidores, know-how y credibilidad.*



Fuente: Adaptado de More than meets the eye: The functional components underlying influencer marketing, por C. Campbell y J. R. Farrell, 2020, Business Horizons, 63(4), p. 3.

El marketing de afiliación surge en paralelo al marketing de influencers. Es un modelo a través del cual las empresas contratan y compensan a los influencers que atraen prospectos o

generan tráfico a sus productos. Con este tipo de acuerdo, los influencers actúan como intermediario entre la empresa y sus seguidores (Dajah, 2020). Al adquirir los servicios de los influencers, las empresas persiguen también adquirir a su audiencia objetivo.

2.1.2. Análisis churn en la era digital

El nacimiento de Internet y el desarrollo de las tecnologías de la información en las últimas décadas ha tenido un gran impacto en la vida de los consumidores y las empresas. Estas últimas han aprovechado esta expansión para diversificar sus fuentes de ingresos y abrirse camino en nuevos mercados, lo que ha contribuido al crecimiento exponencial del comercio electrónico (Shobana et al., 2023). Esto ha generado un ambiente de alta competitividad, ya que los consumidores cuentan con una amplia gama de productos con características y precios similares. Como resultado, la lealtad de los clientes a largo plazo ha disminuido y su permanencia en la misma empresa se ha tornado volátil.

Los elevados recursos que las empresas destinan a intensificar la promoción de sus productos demuestran que su supervivencia depende en gran medida de su capacidad para retener a los clientes y mantenerlos activos. En esencia, existen cuatro tipos de cliente: nuevo, activo, inactivo y perdido. El coste de adquirir a los primeros es elevado, ya que se precisan varias transacciones por parte del nuevo cliente para cubrir los gastos de inversión en su adquisición (Shobana et al., 2023).

Es tarea de cada empresa definir cuándo considera que ha perdido a un cliente. Por ejemplo, Peter Fader (2020) argumenta en su libro Customer Centricity que los clientes que han comprado una gran cantidad de productos de forma reciente son clientes activos y con un mayor potencial futuro. En cambio, aquellos que han realizado alguna compra esporádica, pero no reciente, no se habrían dado necesariamente de baja, sino que podrían ser clientes inactivos. Por último, aquellos que han realizado una gran cantidad de compras, pero no de forma reciente, probablemente están perdidos.

La pérdida de clientes se define de diferentes maneras según el contexto de la industria. Sin embargo, generalmente describe el momento en que un cliente decide terminar su relación con una empresa, ya sea cancelando una suscripción, rescindiendo un contrato o dejando de realizar compras. Además, hay dos tipos de abandono de clientes, el abandono contractual y el abandono no contractual.

1. Clientes con una relación contractual: las transacciones entre los clientes y la empresa están estipuladas por contrato de forma vinculante, de modo que el cliente incurre en un coste más elevado por abandonar. Esto se da sobre todo en las compañías de seguros o telefonía (Xia & He, 2018).
2. Clientes con una relación no contractual: la relación entre el cliente y la empresa se inicia con la primera transacción, sin necesidad de firmar un contrato. En este caso, el cliente no incurre en ningún coste por abandonar; por ello, la tasa de churn en este contexto suele ser mayor (Xia & He, 2018). Este es el caso de los influencers de stylink.

En la actualidad, existe una amplia literatura sobre la predicción de la pérdida de clientes, tanto en contextos contractuales como no contractuales. La mayoría se centra en probar y comparar diferentes algoritmos individuales o en conjunto para optimizar el rendimiento del modelo y reducir el error de predicción (Kim y Lee, 2021). Sin embargo, son pocos los estudios que se centran en su aplicación práctica en el marketing de afiliación e influencers.

Para predecir la tasa de abandono y ayudar a las empresas a optimizar sus estrategias de retención, en la mayoría de los casos se utiliza el aprendizaje supervisado, concretamente las técnicas de clasificación (Misirlis y Vlachopoulou, 2021). Existen numerosos algoritmos que se pueden aplicar a este problema. A continuación, se revisan algunos ejemplos.

En primer lugar, Hassouna et al. (2015) exploran las posibilidades de aplicar modelos de regresión logística y árboles de decisión a la predicción churn de los clientes de una compañía telefónica británica. Los árboles de decisión obtienen un mejor rendimiento.

Vafeiadis et al. (2015) utilizan una base de datos sintéticos pública con clientes del sector de las telecomunicaciones. Comparan los modelos más populares en el momento y el resultado con una mejor métrica de precisión lo obtienen los modelos en los que se aplica boosting, lo que parece dar indicios de que los algoritmos de aprendizaje en conjunto obtienen un mejor rendimiento que los individuales.

Mishra y Reddy (2017) utilizan una base de datos pública para llevar a cabo un análisis comparativo de churn en la industria de las telecomunicaciones. Aplican modelos de aprendizaje en conjunto y el bosque aleatorio obtiene los mejores resultados en cuanto a precisión, error más bajo, menor especificidad y mayor sensibilidad.

Höppner et al. (2017) introducen una nueva técnica, llamada "ProfTree", que es una variante del árbol de decisión clásico, y una nueva métrica de rendimiento llamada "EMPC"[1], que permite identificar el modelo más rentable económicamente para la empresa. Se obtiene que el modelo ProfTree presenta un mejor rendimiento en cuanto a su sensibilidad y maximización de beneficios, pero funciona peor que los otros en cuanto a las métricas AUC[2] y MER[3].

Yanfang y Chen (2017) intentan predecir el abandono de un conjunto de clientes en el ámbito del comercio digital utilizando la técnica de regresión logística. El modelo obtiene una exactitud del 85%.

Rachid et al. (2018) también se centran en el análisis churn en el contexto del comercio electrónico, en este caso con una base de datos de clientes procedentes de una empresa marroquí. Su estudio combina la clusterización de k-medias con LRFM[4], seguida de modelos de aprendizaje supervisado. De nuevo, el modelo de árbol de decisión en conjunto funciona mejor a la hora de identificar a los clientes churn.

Sabbeh (2018) realiza un extenso estudio comparativo donde aplica diez modelos de predicción churn a una base de datos perteneciente a una compañía de telecomunicaciones. Los modelos con mejor rendimiento son, una vez más, los de aprendizaje en conjunto, bosque aleatorio y AdaBoost.

Xia y He (2018) utilizan los datos transaccionales de los clientes de una web china para probar la eficiencia de las redes neuronales de retro propagación y las máquinas de soporte vectorial. Los resultados obtenidos indican, una vez más, que los modelos funcionan mejor en conjunto.

Raeisi y Sajedi (2020) exploran el uso de árboles de decisión potenciados por gradiente (GBT) para predecir el abandono de clientes de la empresa de entrega de comida a domicilio más grande de Teherán (Irán). Para comparar su rendimiento, aplican también otras cinco técnicas de aprendizaje supervisado al conjunto de datos. La mejor exactitud la obtiene el modelo GBT.

[1] Expected maximum profit measure for customer churn (beneficio máximo esperado del abandono de los clientes).
[2] Area Under The Curve (Área bajo la curva) = TP/FP (Verdadero Positivo/Falso Positivo).
[3] Misclassification Error Rate (Tasa de error de clasificación) = Número de clasificaciones incorrectas/Total observaciones.
[4] Length-Recency-Frequency-Monetary.

Por último, en uno de los estudios más recientes, Kim y Lee (2021) analizan los datos de una agencia de marketing de influencers coreana para predecir la tasa de abandono de los seguidores de los influencers a través de un modelo de árbol de decisión. La métrica de exactitud es de un 82%. Se presentan los resultados de forma detallada en la Tabla 1.

Tabla 1. *Revisión de la literatura sobre modelos de predicción churn.*

Autor/es	Marco temporal	Definición churn, sector	Modelos comparados	Mejor modelo
• Hassouna et al. (2015)	• Duración de los contratos, de 12 o de 18 meses	• Telefonía, cliente no renueva contrato de teléfono después de los 12 o 18 meses	• Árbol de decisión • Regresión logística	• Árbol de decisión
• Vafeiadis et al. (2015)	• [N.D.]	• Telecomunicaciones, [N.D.]	• Redes Neuronales Artificiales de múltiples capas, Árboles de Decisión, Máquinas de Soporte Vectorial, Bayes ingenuo y Regresión Logística, Boosting	• Máquina de Soporte Vectorial potenciado con el algoritmo AdaBoost
• Mishra y Reddy (2017)	• [N.D.]	• Telecomunicaciones, [N.D.]	• Bagging , Boosting, árbol de decisión, bosque aleatorio, máquinas de soporte vectorial, y modelo de Bayes ingenuo	• Bosque aleatorio
• Höppner et al. (2017)	• [N.D.]	• Telecomunicaciones, [N.D.]	• ProfTree, árboles de decisión	• ProfTree para maximización de beneficios • Árboles de decisión para AUC y MER
• Yanfang y Chen (2017)	• [N.D.]	• Comercio digital, sin transacciones en los últimos tres meses	• Regresión logística	• Regresión logística 85% precisión
• Rachid et al. (2018)	• Noviembre de 2013 a febrero de 2015	• Comercio electrónico, clientes perdidos según modelo LRFM	• Árbol de decisión, redes neuronales artificiales y árboles de decisión en conjunto	• Árbol de decisión en conjunto
• Sabbeh (2018)	• [N.D.]	• Telecomunicaciones, [N.D.]	• Regresión logística, CART, Bayes ingenuo, Máquinas de soporte vectorial, k vecinos más cercanos, AdaBoost, Descenso de Gradiente Estocástico, Bosque Aleatorio, Perceptrón multicapa, Análisis Discriminante Lineal	• Bosque Aleatorio
• Xia y He (2018)	• Enero a diciembre del 2014	• Website, ninguna transacción durante periodo de tiempo analizado	• Redes neuronales de retropropagación, máquinas de soporte vectorial, combinación	• Modelos en combinación
• Raeisi y Sajedi (2020)	• [N.D.]	• Comercio digital, comida a domicilio, ningún pedido desde hace más de 6 meses	• Árboles de decisión potenciados por gradiente (GBT), K vecinos más cercanos, Bayes ingenuo, Árbol de decisión, Bosque aleatorio, Inducción de reglas	• Árboles de decisión potenciados por gradiente (GBT)
• Kim y Lee (2021)	• Agosto de 2018 y octubre de 2020	• Agencia Marketing, 1 compra de productos promocionados por influencers	• Árbol de decisión	• Árbol de decisión, 82% precisión

Fuente: elaboración propia.

2.1.3. Evaluación del entorno actual: Análisis churn aplicado a stylink

stylink es una empresa alemana de tamaño medio cuyo modelo de negocio se centra en el marketing de afiliación y de influencers[5]. Está presente a nivel internacional en 13 países[6], y es líder de mercado en la región DACH. En 2025, la comunidad de stylink está formada por más de 270.000 influencers y 2.500 marcas asociadas. Al actuar como intermediaria entre influencers y marcas, el modelo de negocio de la empresa se enmarca tanto en un contexto B2B (stylink <-> marcas, redes de afiliación) como B2C (stylink <-> influencers).

El punto de encuentro entre influencers y marcas tiene lugar en la plataforma creada por stylink llamada Linkmaker. Los influencers, en su mayoría nano- y microinfluencers, pueden registrarse a ella de forma gratuita. stylink accede a las cooperaciones con sus marcas asociadas a través de diversas redes de afiliación que, a su vez, son empresas que actúan de intermediarias entre los llamados promotores (stylink) y anunciantes (marcas).

La relación comercial existente entre todos estos partidos se expone a continuación:

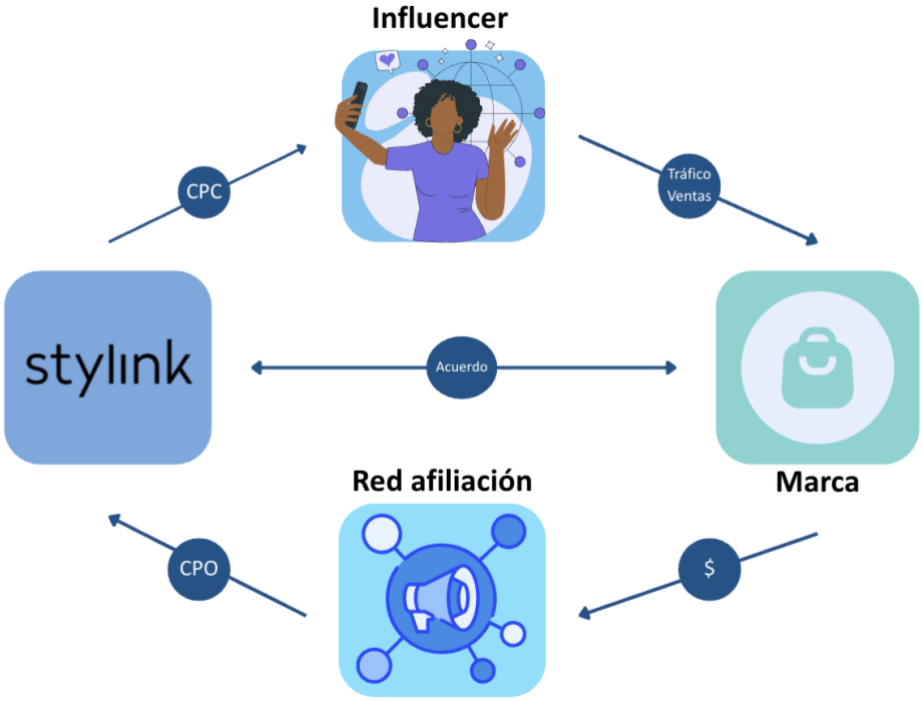
- Las redes de afiliación colaboran con marcas de diversos sectores: moda, alimentación, decoración, deporte, finanzas, entre otras. Algunos ejemplos destacados son: H&M, Amazon, Nike o Sephora. El objetivo de las redes es conectar a dichas marcas con promotores, en este caso stylink, que les ayuden a dirigir más tráfico y generar ventas en sus páginas web.
- stylink se registra como promotor en las diferentes redes de afiliación y, de forma bilateral, acuerda con qué marcas desea asociarse. Por cada venta de productos generada por stylink, la empresa recibe un porcentaje de la transacción, llamado CPO, en concepto de comisión. Una vez cerrado el acuerdo inicial de cooperación, la marca se incluye en la lista que aparece en la plataforma Linkmaker.
- stylink pone a disposición de los influencers de forma gratuita la plataforma Linkmaker. Estos, que pasan a llamarse usuarios después de su registro, pueden elegir cualquiera

[5] Para más información, visite su sitio web: https://www.stylink.com/
[6] En julio de 2025: Alemania, Suiza, Austria, Holanda, Bélgica, Luxemburgo, Francia, Suecia, Polonia, Reino Unido, Irlanda, Estados Unidos y Australia.

de las marcas asociadas presentes en la lista para recomendar sus productos en redes sociales a través de links de afiliación.

- Los influencers copian la URL de los productos que desean recomendar y los transforman en links de afiliación a través de la plataforma Linkmaker. Una vez compartidos en sus redes sociales, estos empiezan a generar una compensación, llamada CPC, por cada clic que reciben por parte de sus seguidores, sin importar si estos compran o no los productos. Esta compensación depende de diversos factores como, por ejemplo, las condiciones del acuerdo firmado entre stylink y la marca correspondiente.

Figura 3. *Modelo de negocio de stylink.*



Fuente: elaboración propia.

stylink ha sido una de las empresas pioneras en escalar este tipo de modelo de negocio. De ahí radica el crecimiento exponencial experimentado en sus primeros años de existencia y su actual liderazgo en el mercado alemán. Sin embargo, al aumentar la popularidad de este, también se ha incrementado la competencia, que ha provocado la elevada tasa de abandono de influencers, especialmente en los últimos dos años.

Cabe señalar que los influencers no ceden su contenido a stylink al compartir sus enlaces de afiliados generados en el Linkmaker, sino que sólo hacen uso del servicio prestado por la

empresa. Se trata de una relación no contractual, lo que lleva implícito un mayor riesgo de abandono.

Con una base de datos y una comunidad de influencers de semejante tamaño, stylink está experimentando dificultades para identificar a los influencers que han abandonado, así como a la hora de clasificarlos según su potencial churn.

2.1.4. Información deseada

El presente trabajo pretende profundizar en la problemática de la elevada tasa de abandono de influencers que usan la plataforma de stylink para monetizar su contenido en redes sociales.

El objetivo es identificar variables que logren explicar los factores que influyen en el abandono para utilizarlas en los modelos que se exponen en apartados posteriores, así como elegir el que obtenga el mejor rendimiento y la mejor métrica de exactitud, lo que permitirá predecir y clasificar a los influencers según su riesgo de churn.

Se observa que el comportamiento en la creación de enlaces y de abandono de los influencers es esencialmente similar al de compra y abandono de los clientes en el proceso transaccional del comercio digital, por lo que es factible utilizar modelos de clasificación para pronosticar el comportamiento de los influencers de stylink.

2.1.5. Información actual: deficiencias y soluciones alternativas.

Para definir los obstáculos que impiden obtener la información planteada en el apartado anterior, es recomendable realizar un análisis que muestre la situación actual de stylink en materia de datos.

El análisis DAFO es una herramienta útil para evaluar a una empresa, un plan, un proyecto o una actividad comercial y definir cuáles son los recursos disponibles, las deficiencias, las oportunidades y las amenazas externas a las que se enfrenta (Gürel & Tat, 2017).

En la Figura 4 se presenta un modelo DAFO en el que se analizan estos factores. Las principales deficiencias, o debilidades, radican en la fase de transición hacia una estrategia basada en datos en la que se encuentra la empresa. Las redes de afiliación, además, no siempre proporcionan toda la información deseada con respecto a las transacciones de los influencers,

y aún no se ha creado una estrategia homogénea y coordinada en todos los mercados para la retención y mejora de la experiencia de cliente.

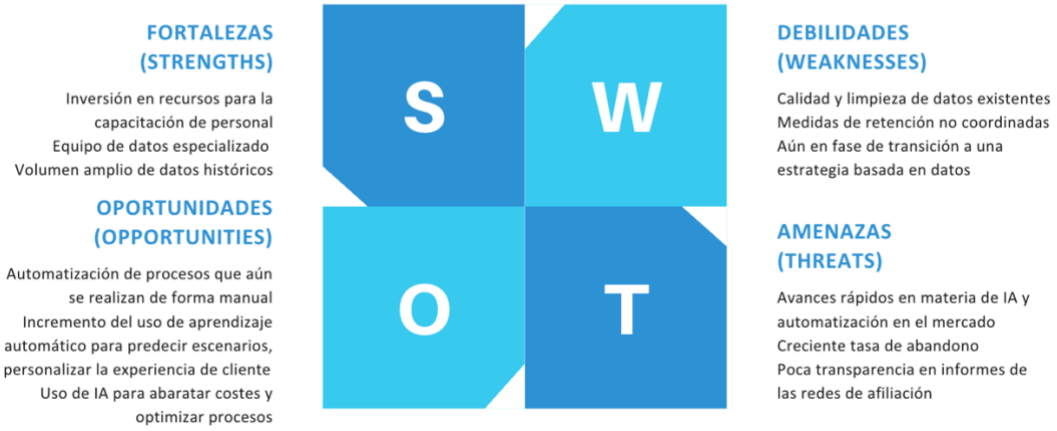
Se deberá, por tanto, realizar una limpieza y preprocesamiento de los datos antes de poder realizar el modelado y proponer una estrategia que se pueda escalar a todos los mercados internacionales en los que opera la empresa.

Este desafío se ve intensificado por las amenazas externas, que consisten fundamentalmente en el rápido avance de la competencia en materia de automatización e implementación de IA en sus procesos, así como la creciente tasa de abandono de influencers.

Sin embargo, de esta situación de amenaza también se derivan oportunidades, y las debilidades se pueden contrarrestar con las fortalezas de la empresa.

Estas son, principalmente, la elevada inversión en capacitación del personal, la contratación de expertos en materia de IA y automatización, la amplia experiencia de stylink en el sector y, por ende, el amplio volumen de datos históricos de los que se dispone.

Figura 4. *Modelo DAFO (SWOT por sus siglas en inglés) aplicado a stylink en materia de datos.*



Fuente: elaboración propia.

2.1.6. Habilidades analíticas actuales.

Desde su fundación y hasta mediados del año 2023, la mayor parte de las decisiones empresariales de stylink se basaron en un modelo HIPPO (Kaushik, 2009). Este es el acrónimo de “Highest Paid Person’s Opinion”, es decir, que las decisiones estaban fundamentalmente basadas en la opinión de la persona mejor pagada en lugar de en datos objetivos.

Este enfoque cambia cuando la empresa crece y la complejidad de los datos y estructuras aumenta. La nueva estrategia prioriza el análisis y la visualización más eficientes de los datos históricos, la reducción de costos y la optimización de ingresos.

Como resultado, las capacidades analíticas de stylink han mejorado significativamente. De cara al futuro, la compañía planea invertir más en automatización, ciencia de datos e inteligencia artificial, áreas clave que respaldarán la creciente demanda e impulsarán la expansión del mercado.

2.2. Fase de análisis: identificación de *gaps*

2.2.1. Capacidad de los informes actuales

Los informes actuales se centran fundamentalmente en el análisis descriptivo y de diagnóstico de los datos históricos de la empresa. Es decir, se busca responder a las preguntas de "qué ha pasado" y "por qué" (Jaramillo-Chuqui & Villarroel-Molina, 2023).

2.2.1.1. Informes transaccionales

stylink recopila información sobre los influencers y su rendimiento generado a través de los enlaces de afiliación que comparten. Se les identifica con un ID de usuario único. Las principales variables que se incluyen en estos informes se pueden dividir en cuatro categorías principales:

- Ingresos:
    - Provisión: cantidad generada por las transacciones realizadas por los seguidores de los influencers al comprar los productos que estos promocionan con sus links de afiliación.
    - Beneficio: provisión neta, después de eliminar la cantidad correspondiente a los productos que son devueltos.
    - Factor: Provisión/Comisión. Es la relación que permite monitorizar si los costes están superando a los beneficios.
- Stats: comisión o CPC pagada al influencer por los clics generados en sus links.
- Links:
    - Link ID: identificador único.
    - Número de links: cantidad de links asociados a un influencer.
- Clics: cantidad de clics asociados a un Link ID.

Otra fuente de datos son los informes que recogen información sobre el tráfico y las transacciones generadas para las marcas asociadas con stylink, a las que se identifica con un Source ID único.

- GMV: valor monetario de los productos adquiridos por los seguidores de los influencers a través de sus enlaces (por sus siglas en inglés, Gross Merchandise Value).
- Ventas: número de transacciones asociadas a un Link ID.
- CPO/CPA: porcentaje de comisión que stylink percibe por cada producto vendido.

Además, se consideran las siguientes variables agregadas:

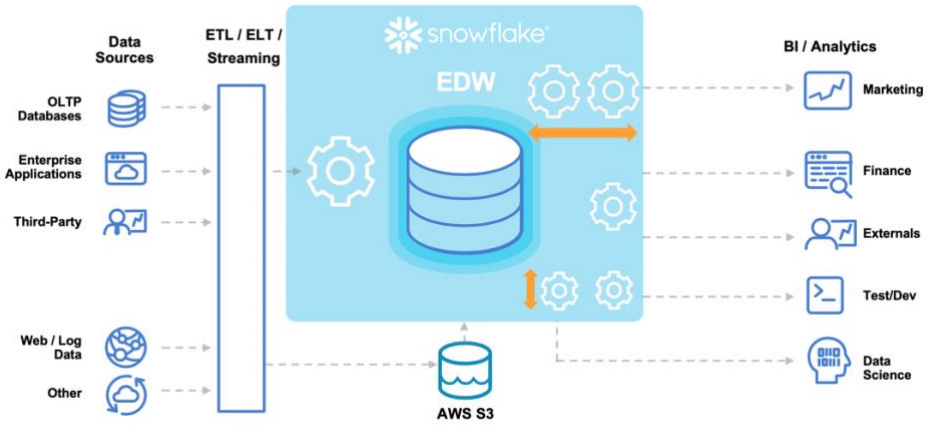
- Diversidad: de cuántas marcas diferentes un influencer crea links, lo que ayuda a determinar la dependencia de stylink a sus marcas asociadas a la hora de mantener a un cliente.
- % Links con transacciones: qué porcentaje de los enlaces creados por los influencers han generado transacciones.
- % Días con transacciones: en qué porcentaje de los días en los que un influencer ha compartido links este ha generado transacciones.
- Días/Meses únicos: días y meses únicos, en valor absoluto, en los que un influencer ha generado transacciones.
- ATV: valor de transacción medio (por sus siglas en inglés, Average Transaction Value), es decir, GMV promedio generado por cada transacción, por influencer.
- Promedio de links por mes: cuántos links al mes ha creado de media un influencer.

2.2.2. Proveedores de tecnología necesarias

Para poder realizar este proyecto, se requiere acceder a las bases de datos de la empresa. stylink hace uso de AWS como proveedor de servicios de computación en la nube para gestionar su infraestructura de datos. Estos, a su vez, proceden de diferentes fuentes.

El equipo de TI se encarga de los procesos ETL necesarios para poder almacenar los datos procedentes de las diferentes fuentes en el almacén de datos Snowflake. Posteriormente, el equipo de datos accede a ellos a través de consultas en lenguaje SQL. Se dispone de todos los datos históricos generados por stylink entre los años 2019 y 2025. Además, se trabaja con el software de visualización Tableau para elaborar informes transaccionales y dashboards que permiten a los steakholders tomar decisiones comerciales estratégicas.

Figura 5. *Sistema de almacenamiento y procesamiento de datos en la nube a través de Snowflake y AWS como proveedor de servicios.*



Fuente: Amazon Web Services (s.f.). Snowflake Partner Solutions for Financial Services. AWS. Recuperado de https://aws.amazon.com/de/financial-services/partner-solutions/snowflake/

2.2.3. Diferencia entre los informes actuales y la información deseada

Los informes actuales, como se ha expuesto en apartados anteriores, se centran en un análisis descriptivo y de diagnóstico de los datos existentes. Sin embargo, el objetivo de este trabajo es aportar un enfoque predictivo y prescriptivo a la hora de evaluar la tasa de abandono de los influencers en stylink; es decir, se pretende responder a las preguntas de "qué va a pasar en el futuro" y "qué se debe hacer" al respecto (Jaramillo-Chuqui & Villarroel-Molina, 2023). En este caso, se pretende predecir qué influencers van a dejar de usar stylink y qué estrategias se deben llevar a cabo para retenerlos antes de que se conviertan en clientes perdidos.

2.2.4. Cronología, costes y recursos humanos implicados

Este trabajo se lleva a cabo entre las semanas del calendario 13 y 27 del año 2025. El proyecto se organiza siguiendo el modelo CRISP-DM (IBM, 2015).

- Primera Fase: Comprensión del negocio.

  En las semanas 13-14, se realiza un análisis del contexto actual en el sector del marketing de afiliación y de influencers. A continuación, se enmarca la presencia de stylink en este contexto. Se realiza una revisión de la literatura existente en materia de análisis churn que sirve de base teórica para el presente trabajo.

A través del análisis DAFO, se presentan los factores internos y externos que contribuyen a la problemática que se pretende resolver, y se define la información existente, la deseada y los gaps existentes entre estas.

Por último, tal y como se detalla en este apartado, se plantean los recursos humanos y tecnológicos necesarios para la consecución de este proyecto y su aplicación práctica.

- Segunda fase: Comprensión de los datos.

En la semana 15, se analizan los datos existentes en las bases de datos de stylink. Con ayuda de los principales stakeholders, se delimita el alcance del análisis a un marco temporal y un mercado determinados. Estos definen también el concepto de churn en el contexto de los influencers de stylink.

Se accede al almacén de datos Snowflake y se extraen los datos necesarios a través de consultas SQL. Se procede al análisis descriptivo de las variables, con objeto de comprender mejor las distribuciones, los patrones y las correlaciones, y a la selección de aquellas que puedan explicar la pérdida de influencers.

- Tercera fase: Preparación de los datos.

Entre las semanas 16-18, se preparan los datos para su posterior modelado. Se realiza una limpieza, que consiste en el tratamiento de los valores nulos, la eliminación de valores atípicos, duplicados, variables redundantes, etc.

Posteriormente, se realiza el preprocesamiento de datos, que incluye la normalización, la creación de nuevas variables en el data set, el escalado, cambios en los tipos de variable, etc.

- Cuarta fase: Modelado de los datos.

En las semanas 19-21, con el conjunto de datos limpio y preprocesado, se realiza el propio modelado. Se dividen los datos en subconjuntos de entrenamiento y prueba para su posterior evaluación. Se aplican los dos modelos elegidos, regresión logística y árbol de decisión, y se validan los resultados.

- Quinta fase: Evaluación de los modelos.

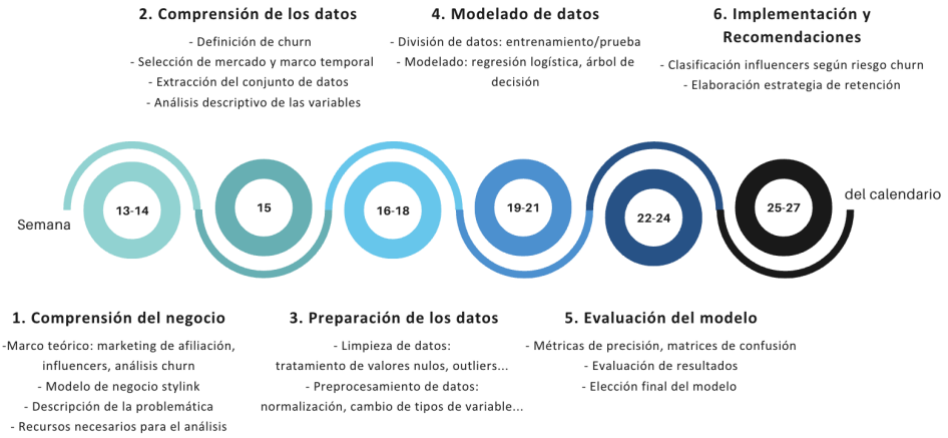
En las semanas 22-24, se utilizan diferentes métricas de validación, como matrices de confusión, para evaluar la calidad de los modelos. Se analizan los resultados y se

comprueba si estos identifican con éxito a los influencers con un riesgo de abandono más elevado. Se elige, además, el que obtiene un mejor rendimiento.

- Sexta fase: Implementación y recomendaciones.

En las semanas previas al depósito del trabajo, la 25-27, se divide a los influencers en categorías según su riesgo de abandono y se propone una estrategia de retención que se ajuste a las necesidades de stylink.

Figura 6. *Cronología de proyecto basada en el modelo CRISP-DM.*



Fuente: elaboración propia.

Los costes de realizar este proyecto están financiados íntegramente por stylink. Incluyen el uso de todo tipo de software y hardware, materiales, bases y almacenes de datos, herramientas de visualización y personal del equipo de Inteligencia de Negocio.

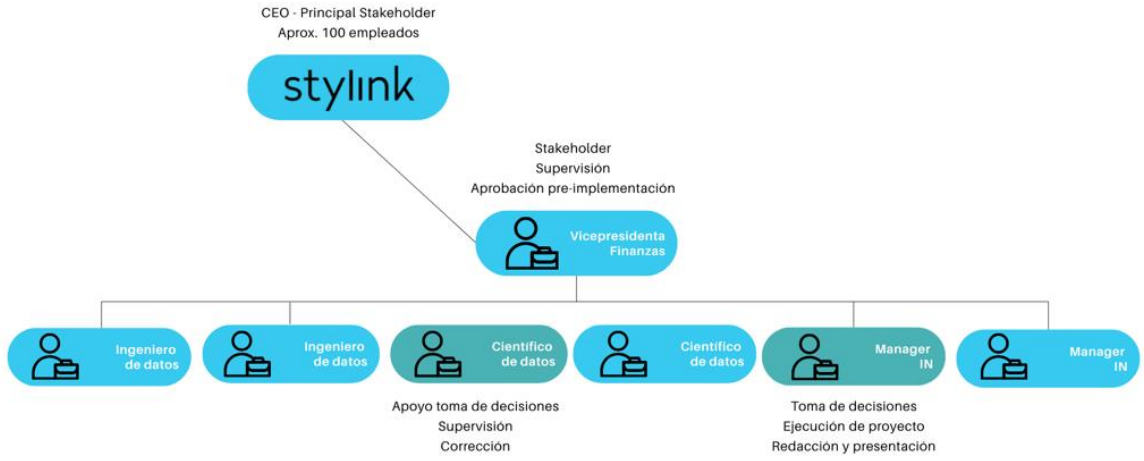
Figura 7. *Costes de proyecto.*

Fuente: elaboración propia.

Este proyecto tiene como principales stakeholders al CEO de stylink, que representa los intereses de la empresa, así como a la vicepresidenta de Finanzas. Esta última se encarga de supervisar el proyecto y aprobar los resultados, así como de proporcionar sugerencias estratégicas, como paso previo a la fase de implementación.

Los recursos humanos implicados en la consecución del proyecto consisten en una científica de datos, que forma parte del equipo de datos, así como una mánager de inteligencia de negocio, autora del presente trabajo.

Figura 8. *Recursos humanos implicados en el proyecto y roles.*



Fuente: elaboración propia.

La primera se encarga de apoyar en la toma de decisiones, por ejemplo, durante la fase de modelado, así como de supervisar y revisar los resultados. La segunda se encarga de la elaboración, ejecución, redacción, presentación y puesta en marcha del proyecto.

La carga de trabajo para la científica de datos se incluye en sus 40 horas semanales estipuladas por contrato. En el caso de la mánager de inteligencia de negocio, se extienden a 4 horas adicionales diarias.

2.3. Fase de recomendaciones: alcance, prioridades y presupuesto

2.3.1. Promoción del proyecto en la organización

El proyecto se presenta al equipo de datos y a los principales stakeholders en la semana 25, lo que da inicio a la sexta fase del proyecto.

Una vez aprobado, se implementa el modelo en el mercado seleccionado y se presenta al equipo de mercados internacionales la estrategia de retención a seguir en semanas y meses posteriores.

Finalmente, se elaboran las conclusiones pertinentes, se analizan las limitaciones encontradas y se concretan los pasos a seguir en estudios posteriores.

No hay un límite presupuestario establecido y los costes incluyen el uso de software, hardware, licencias, materiales y los recursos humanos resumidos en las Figuras 7 y 8. Este presupuesto asciende a un total aproximado de 2.000€ durante los 5 meses de duración del proyecto.

3. Análisis y definición

3.1. Análisis de los datos a utilizar

3.1.1. Comprensión de los datos

Como se ha mencionado en apartados anteriores, la segunda fase del proyecto según el modelo CRISP-DM consiste en la comprensión de los datos (IBM, 2015).

El conjunto de datos incluye enlaces de afiliación creados por influencers alemanes, así como los datos transaccionales asociados a estos, en un periodo que abarca desde enero de 2023 hasta abril de 2025. El análisis comienza en el punto en el que se observa por primera vez un aumento significativo del churn y continúa hasta la fecha de extracción de los datos.

El siguiente paso consiste en determinar los criterios que describen a un influencer “perdido” para la empresa. Puesto que la relación de los influencers con stylink no es contractual, el registro en la plataforma no implica automáticamente un uso activo: un influencer puede simplemente crear una cuenta sin generar nunca enlaces. Asimismo, no todos los enlaces dan lugar a transacciones. Además, de acuerdo con el modelo de negocio de la empresa, se compensa a los usuarios por clic y no por transacción.

Por otra parte, no todos los usuarios que comparten sus enlaces aportan un valor monetario a stylink. Un usuario solo se considera activo una vez que genera su primera transacción o, lo que es lo mismo, provisión. Además, su nivel de actividad puede variar con el tiempo y en función de las marcas que promociona.

Es difícil relacionar directamente la tasa de abandono con la disminución de los enlaces creados. Sin embargo, la pérdida de influencers suele estar relacionada con su última actividad. En el presente trabajo, se clasifican como "perdidos" a aquellos que no han creado un enlace nuevo en 90 o más días.

En resumen, el conjunto de datos a analizar parte de las siguientes premisas:

- Los influencers están registrados en la plataforma de stylink Alemania.
- Han creado enlaces entre el enero de 2023 y abril de 2025.
- Han creado al menos un enlace de afiliación en este periodo de tiempo.
- Uno o más de sus enlaces han generado transacciones y valor monetario, es decir, provisión para la empresa (usuarios activos).
- Se consideran perdidos a aquellos influencers que han creado su último enlace hace 90 o más días.

El objetivo que se persigue en los siguientes apartados es encontrar variables que expliquen la pérdida de clientes, utilizarlas para construir dos modelos de clasificación, comparar su rendimiento, y elegir el que mejor prediga el abandono de los influencers.

3.1.2. Análisis de datos exploratorio (EDA)

El análisis exploratorio constituye un paso fundamental para profundizar en la información contenida en el conjunto de datos. Incluye parámetros como los valores máximos y mínimos, la media, la distribución de los datos, así como otras medidas estadísticas que permiten detectar la presencia de valores atípicos y nulos, entre otros (Sharma, Patel & Shrivastava, 2023).

Los datos extraídos para el presente trabajo reflejan el comportamiento de los influencers con respecto a sus enlaces, provisión, comisión, transacciones generadas, su fecha de registro en la plataforma, su ID identificador, el ID identificador del enlace de afiliación, la fecha en la que este se creó, y el ID de la marca para la que crearon el enlace, desde enero de 2023 hasta abril de 2025.

Los datos están agrupados por ID de enlace. Sin embargo, el análisis churn va a realizarse a nivel de usuario, por lo que el primer paso consiste en crear un nuevo conjunto de datos que agrupe las observaciones por usuario, de modo que solamente exista una fila por cada USER_ID único. Para ello, se toma la suma total para agregar el GMV, la provisión, la comisión,

el número de transacciones, y el número de enlaces. El nuevo conjunto de datos se compone de 7 variables y 15.601 observaciones.

No se detecta la presencia de valores nulos ni de filas que contengan duplicados. Por otro lado, al analizar las características generales de las columnas numéricas, se observa que la columna GMV contiene valores negativos. Estos son datos erróneos que deberán eliminarse en la fase de limpieza.

El tipo de variables de las que se compone el conjunto de datos son numéricas sin decimales en el caso de USER_ID, NO_TRANSACTIONS y LINK_COUNT, object o de texto en el caso de REGISTER_DATE, y float o numéricas con decimales en el caso de GMV, PROVISION y COMMISSION. También se deberán realizar cambios en el tipo de variables.

Tabla 2. *Análisis descriptivo de las variables del conjunto de datos agrupado por USER_ID.*

	USER_ID	GMV	PROVISION	COMMISSION	NO_TRANSACTIONS	LINK_COUNT
count	15601.000000	1.560100e+04	1.560100e+04	15601.000000	15601.000000	15601.000000
mean	161277.019614	5.088109e+04	2.429851e+03	436.582654	511.655727	58.171335
std	73357.699398	4.627572e+05	2.467871e+04	3550.537464	4573.145779	404.606429
min	4.000000	0.000000e+00	0.000000e+00	0.000000	0.000000	1.000000
25%	114912.000000	0.000000e+00	0.000000e+00	0.480000	0.000000	2.000000
50%	171582.000000	4.540000e+01	1.760000e+00	3.240000	1.000000	8.000000
75%	201043.000000	2.263860e+03	1.019700e+02	35.370000	24.000000	29.000000
max	284292.000000	2.281434e+07	1.533669e+06	236597.130000	177876.000000	16247.000000

Fuente: elaboración propia.

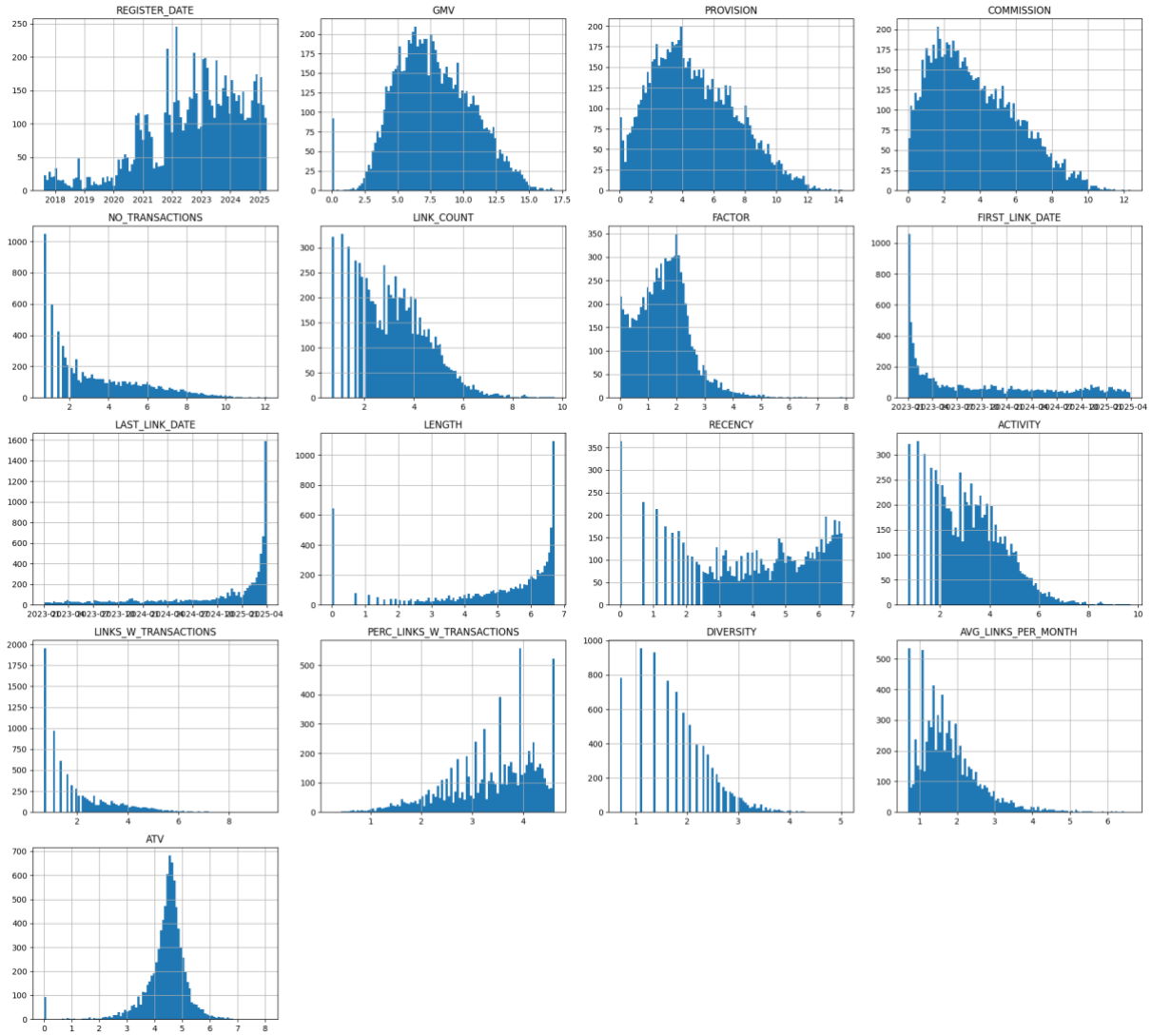
Las estadísticas generales que se muestran en la Tabla 2 indican que, de media, en el periodo observado, los influencers han generado un GMV de 50.881€, una provisión de 2.429€, una comisión de 436€, así como un promedio de 511 transacciones y 58 enlaces. Esta información servirá para comparar al usuario promedio con los que presentan comportamientos atípicos.

Los valores máximos dan indicios de la presencia de outliers. A continuación, se utilizan diagramas de caja para comprobarlo. Se confirma que en todas las variables existen numerosos valores atípicos que se explorarán en el siguiente apartado.

Para analizar la distribución de los datos se crean histogramas de las variables. Se observa en la Figura 9 que gran parte de los USER_ID no ha creado enlaces que hayan generado transacciones, lo que dificulta observar la distribución en el resto de las variables ya que, al crear los gráficos, las categorías están desequilibradas.

Se puede aplicar una transformación logarítmica para normalizar los datos y comprender mejor su distribución. A continuación, se verifica que el número total de usuarios con links sin provisión es de 7.305, lo que supone un 46,8% de la muestra.

Figura 9. *Histogramas de las variables del conjunto de datos de stylink.*



Para poder encontrar las variables que explican el abandono de los influencers, es fundamental analizar la correlación existente entre las mismas. Las matrices de correlación proporcionan información acerca de la relación linear existente entre dos o más variables. El coeficiente de correlación puede variar entre -1 y 1. Cuanto más próximos a estos valores, más fuerte será la correlación, que puede ser negativa o positiva. Una correlación positiva implica que, cuando una variable aumenta, la otra también aumenta. El caso contrario sucede con la correlación negativa. Un coeficiente de correlación igual a 0, por otra parte, indica que no existe ningún tipo de relación entre las variables (Misirlis & Vlachopoulou, 2021).

En este conjunto de datos, se observa una correlación positiva muy fuerte entre la provisión y el GMV (0,85), el GMV y la comisión (0,84), el GMV y el número de transacciones (0,97), la provisión y el número de transacciones (0,90), la provisión y la comisión (0,94), así como la comisión y el número de transacciones (0,87). Además, existe una fuerte correlación positiva entre el número de links creados y el número de transacciones generadas (0,60). Por último, existe una correlación positiva moderada entre el GMV y el número de links (0,53), el número de links y la provisión (0,44) y la comisión y el número de links (0,44).

Como se puede observar, todas las variables están estrechamente relacionadas, lo que podría causar problemas de multicolinealidad. La multicolinealidad significa que dos o más variables transmiten prácticamente la misma información, lo que afecta a la robustez de los modelos. Para evitarla, se pueden utilizar diferentes técnicas que se exploran en la fase de preprocesamiento.

Una vez realizado el análisis descriptivo de las variables principales del conjunto de datos, se procede a su limpieza y preprocesado.

3.2. Análisis histórico y/o limpieza de datos

3.2.1. Preparación de los datos

Según el modelo CRISP-DM, la tercera fase del proyecto consiste en la preparación de los datos (IBM, 2015). La limpieza incluye el tratamiento de los valores anómalos, la eliminación de inconsistencias y valores nulos, así como la creación de nuevas variables. El preprocesamiento, por su parte, comprende la normalización, estandarización, escalado o el cambio de tipo de variables (Misirlis & Vlachopoulou, 2021).

3.2.1.1. Limpieza de los datos

En primer lugar, se eliminan los valores incoherentes. En el apartado anterior, se observó la presencia de GMV negativos que se deben eliminar, ya que el GMV siempre es 0€ o superior. Se comprueba que se trata de un solo valor negativo y se elimina en el conjunto de datos original, de forma que no se incluya en la suma del GMV al agregar a nivel de USER_ID.

En segundo lugar, se eliminan a aquellos USER_ID que no cumplen con los parámetros establecidos en la fase de comprensión de los datos, es decir, a aquellos que solo han creado enlaces cuya provisión es igual a 0€. En la fase de análisis, se encontró que se trata de 7.305

usuarios. Por tanto, el número de USER_ID únicos que han generado enlaces con provisión (usuarios activos) es de 8.296. A continuación, como se muestra en la Tabla 3, se procede a la creación de las nuevas variables.

Tabla 3. *Variables del conjunto de datos después de la eliminación de los usuarios sin provisión y la creación de nuevas variables.*

USER_ID	object
REGISTER_DATE	datetime64[ns]
GMV	float64
PROVISION	float64
COMMISSION	float64
NO_TRANSACTIONS	int64
LINK_COUNT	int64
FACTOR	float64
FIRST_LINK_DATE	datetime64[ns]
LAST_LINK_DATE	datetime64[ns]
LENGTH	int64
RECENCY	int64
ACTIVITY	int64
LINKS_W_TRANSACTIONS	int64
PERC_LINKS_W_TRANSACTIONS	float64
DIVERSITY	int64
AVG_LINKS_PER_MONTH	float64
ATV	float64
CHURN	bool
dtype: object	

Fuente: elaboración propia.

La fecha del primer enlace (FIRST_LINK_DATE), el último enlace (LAST_LINK_DATE) y la longitud (LENGTH), es decir, la diferencia de días entre el último y primer link, se crean porque podrían dar indicios sobre el comportamiento de los influencers que abandonan. Se puede suponer que la tendencia al abandono es mayor en influencers que dejan transcurrir más tiempo entre enlaces.

La actividad (ACTIVITY) indica el número de días únicos en el que los usuarios han creado sus enlaces. Además, si se calcula cuántos han creado links solamente en un día (ONE_DAY) estos son solamente 322 influencers, lo que supone un 3,8% de todos los usuarios activos del conjunto de datos. Esto implica que los influencers que superan la barrera de crear su primer enlace suelen repetir este comportamiento.

Con esta variable se pretende identificar si los usuarios que tienen una menor actividad o son usuarios de un día tienen una mayor probabilidad de abandono.

La recencia (RECENCY) proporciona información sobre el número de días transcurridos entre la fecha de extracción de los datos y el último link creado por el influencer. En este caso, se pretende confirmar si los usuarios menos recientes tienen un mayor riesgo de churn.

LINK_COUNT refleja el número total de enlaces creados por los influencers durante el periodo observado. El número de links (LINKS_W_TRANSACTIONS) y el porcentaje de links con transacciones (PERC_LINKS_W_TRANSACTIONS) proporcionan información sobre la rentabilidad de los enlaces de los influencers. Se busca comprender si el número total de enlaces, con y sin transacciones, influyen en la tasa de abandono.

La diversidad (DIVERSITY) indica el número único de marcas (SOURCE_ID) para las que los influencers han creado enlaces. Podría suceder que los usuarios con una menor diversidad en sus enlaces tuvieran un mayor riesgo de abandono ya que, en caso de que stylink terminara su cooperación con dichas marcas, se perdería automáticamente a los influencers también.

El promedio de links que los usuarios crean al mes (AVG_LINKS_PER_MONTH) es también una métrica que podría dar indicios de su riesgo de churn. Aquellos que son menos activos mensualmente podrían ser más propensos a abandonar.

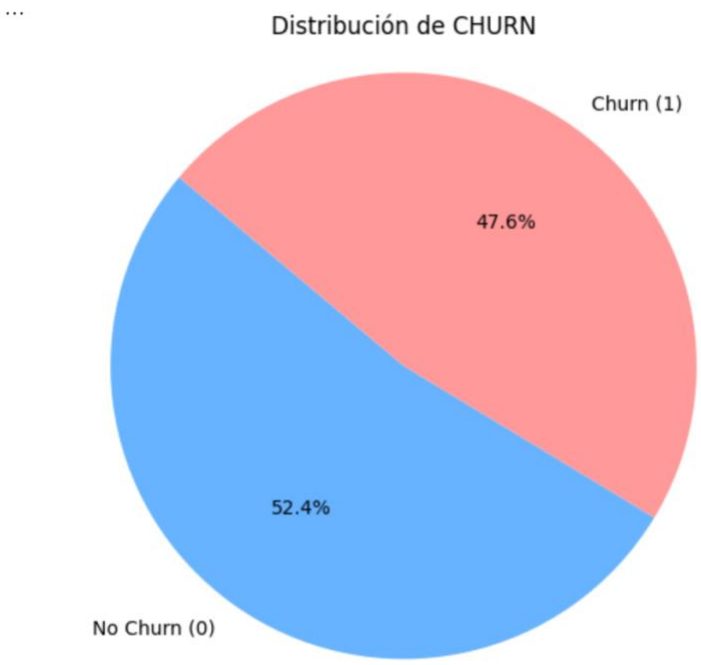
El valor medio de las transacciones de los usuarios (ATV) se calcula dividiendo el GMV total generado por los enlaces del USER_ID entre el número de transacciones. Esta métrica también proporciona información acerca de la rentabilidad de los usuarios y podría ser indicativa de la tasa de abandono, si los usuarios que generan un GMV más bajo con sus enlaces fueran aquellos que más tienden a abandonar.

Se calcula la variable FACTOR, un KPI relevante que la empresa utiliza para medir el rendimiento de los enlaces a nivel de usuario y de mercado. Es el resultado de dividir la provisión total de todos los enlaces de un usuario entre la comisión total que percibe de sus clics.

A nivel de mercado, se calcula dividiendo la provisión total generada por todos los influencers de dicho mercado entre la comisión total percibida por los mismos. La razón para incluir esta métrica es identificar patrones en el rendimiento económico de los usuarios perdidos.

Por último, se crea una columna CHURN para analizar cuántos influencers han abandonado. (0) significa no churn y (1) churn. Se observa en la Figura 10 que, de los 8.296 influencers, el 52,45% no ha abandonado y el 47,55% sí ha abandonado. En valores absolutos, esto significa que 4.351 no han abandonado y 3.945 sí han abandonado.

Figura 10. *Distribución de los influencers de stylink Alemania en función del churn.*



Fuente: elaboración propia.

El nuevo conjunto de datos está compuesto por 19 variables y 8.296 observaciones. Se comprueba que no haya valores nulos ni duplicados. Además, se analiza de nuevo la distribución (Figura 9), y se observa que solo las variables GMV, PROVISION, COMMISSION y ATV poseen una distribución relativamente normal, lo que hará necesario una estandarización en la fase de procesamiento para construir modelos sin sobreajuste.

El siguiente paso consiste en analizar los outliers. Para ello, se crean diagramas de caja de las variables numéricas. En la Figura 11, se aprecian los numerosos valores atípicos que se dan en todas las variables. Esto indica que el comportamiento de una gran parte de los influencers no sigue un patrón regular, lo que dificultará el entrenamiento de los modelos.

A continuación, en la Tabla 4, se muestran los valores obtenidos en los diagramas de caja de la Figura 11 para comprender el alcance de estos usuarios con comportamientos anómalos. El

análisis indica que 3.343 USER_ID únicos y 11.199 filas presentan valores anómalos en varias de sus métricas. Esto supone un 40,30% del conjunto de datos.

En algunos casos, los valores atípicos se eliminan o se limitan a un valor determinado según el caso de estudio. Sin embargo, no siempre es recomendable eliminarlos, ya que pueden proporcionar mucha información oculta en los datos (Misirlis & Vlachopoulou, 2021).

Al explorar en profundidad cuáles son las variables que presentan más valores anómalos, se obtienen los resultados presentados en la Tabla 4. Las variables monetarias, es decir, GMV, número de transacciones, provisión y comisión, son las que más destacan. Esto puede deberse a la popularidad y explosión en ventas de algunos productos que los influencers recomiendan, ya sea por estar estos en tendencia, o por el poder de venta del propio influencer, en comparación con el ATV del resto de sus enlaces, que se acerca más a la media.

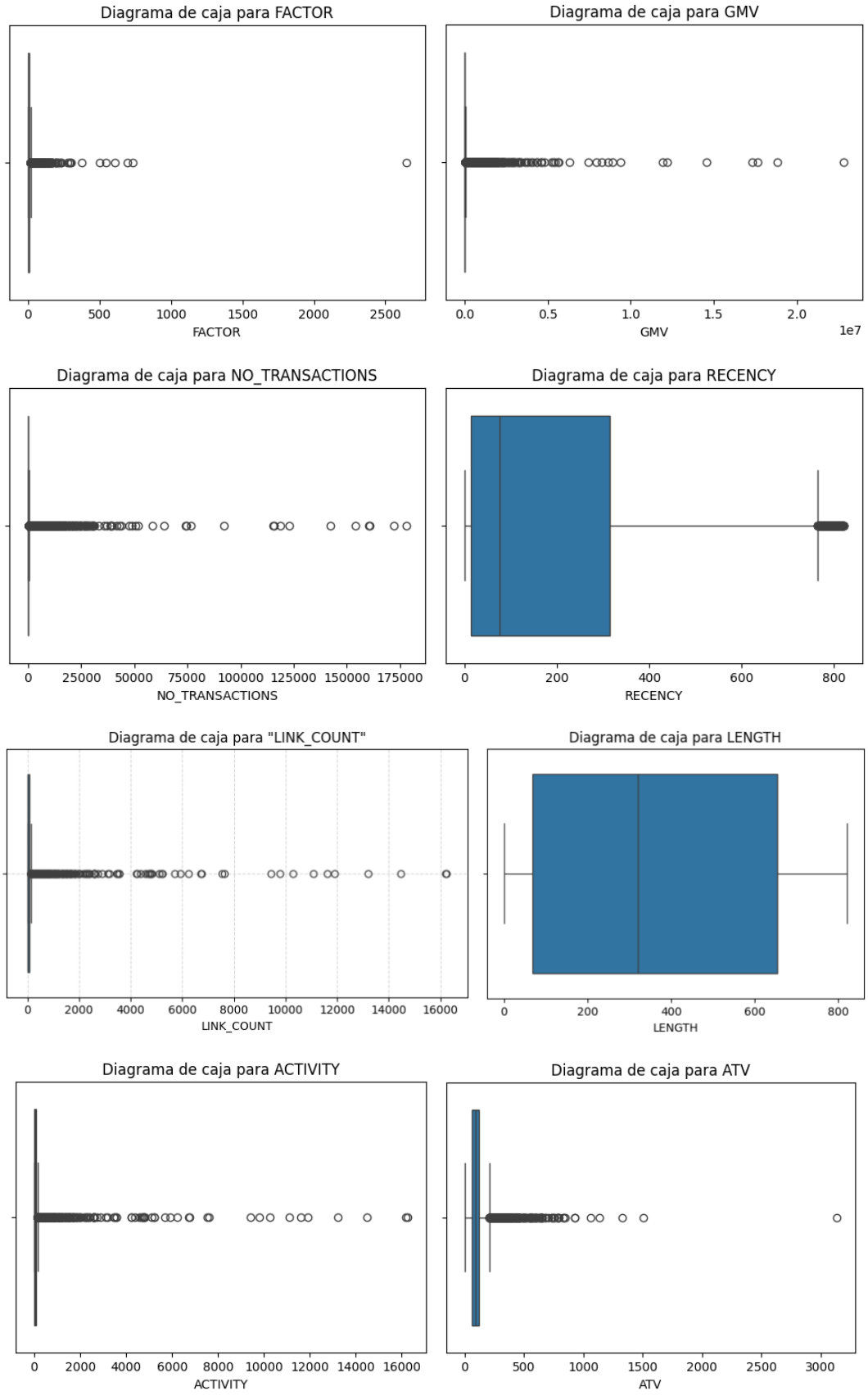
Tabla 4. *USER_ID y ránking de variables que presentan más valores atípicos.*

```
[11199 rows x 3 columns]
Usuarios únicos con outliers: 3343
Columna
GMV                    1437
NO_TRANSACTIONS        1381
PROVISION              1378
COMMISSION             1359
LINKS_W_TRANSACTIONS   1140
LINK_COUNT              950
ACTIVITY                950
AVG_LINKS_PER_MONTH     828
FACTOR                  616
DIVERSITY               541
ATV                     454
RECENCY                 165
Name: count, dtype: int64
```

Fuente: elaboración propia.

El tratamiento de los valores atípicos se ve influenciado por el contexto de uso empresarial. En algunos casos, es necesario eliminarlos por completo para no distorsionar los resultados del análisis. En otros, como el presente, se opta por mantenerlos para evitar prescindir de un 40% de información necesaria y relevante para predecir el abandono de los influencers, ya que los usuarios con valores extremadamente altos en las variables monetarias suelen ser aquellos con mayor valor estratégico para stylink.

Figura 11. *Diagrama de cajas de las variables numéricas para detectar valores atípicos.*



Fuente: elaboración propia.

De nuevo, se analizan las relaciones que se dan entre las nuevas variables del conjunto de datos. Además de la fuerte correlación existente entre las métricas monetarias, cabe destacar que existe una fuerte correlación positiva entre CHURN y RECENCY (0,76) y una correlación negativa moderada entre CHURN y LENGTH (-0,44). Podría darse que, cuanto más tiempo transcurre entre el primer y el último link, siempre que el influencer continúe publicando enlaces de forma regular, más disminuye la posibilidad de abandono.

Otras correlaciones relevantes se pueden observar en la Figura 12 y se dan entre LINKS_W_TRANSACTIONS y LINK_COUNT (0,95), AVG_LINKS_PER_MONTH, ACTIVITY y LINK_COUNT (0,90), LINKS_W_TRANSACTIONS Y AVG_LINKS_PER_MONTH (0,83), NO_TRANSACTIONS y LINK_COUNT (0,60), RECENCY y LENGTH (-0,53), así como GMV y LINKS_W_TRANSACTIONS (0,60), LINK_COUNT y GMV (0,53).

Es intuitivo pensar que, cuando un influencer publica un elevado número de enlaces, su probabilidad de generar transacciones será mayor y, por consiguiente, también aumentará el número total de transacciones, el GMV y la provisión generada. Sin embargo, la experiencia en el negocio demuestra que esta correlación varía enormemente en función de los usuarios.

Además, cuando aumenta el número de enlaces (con transacciones), también aumenta la media de enlaces mensual. Si esta aumenta, en consecuencia, también aumentará su actividad y su número total de enlaces.

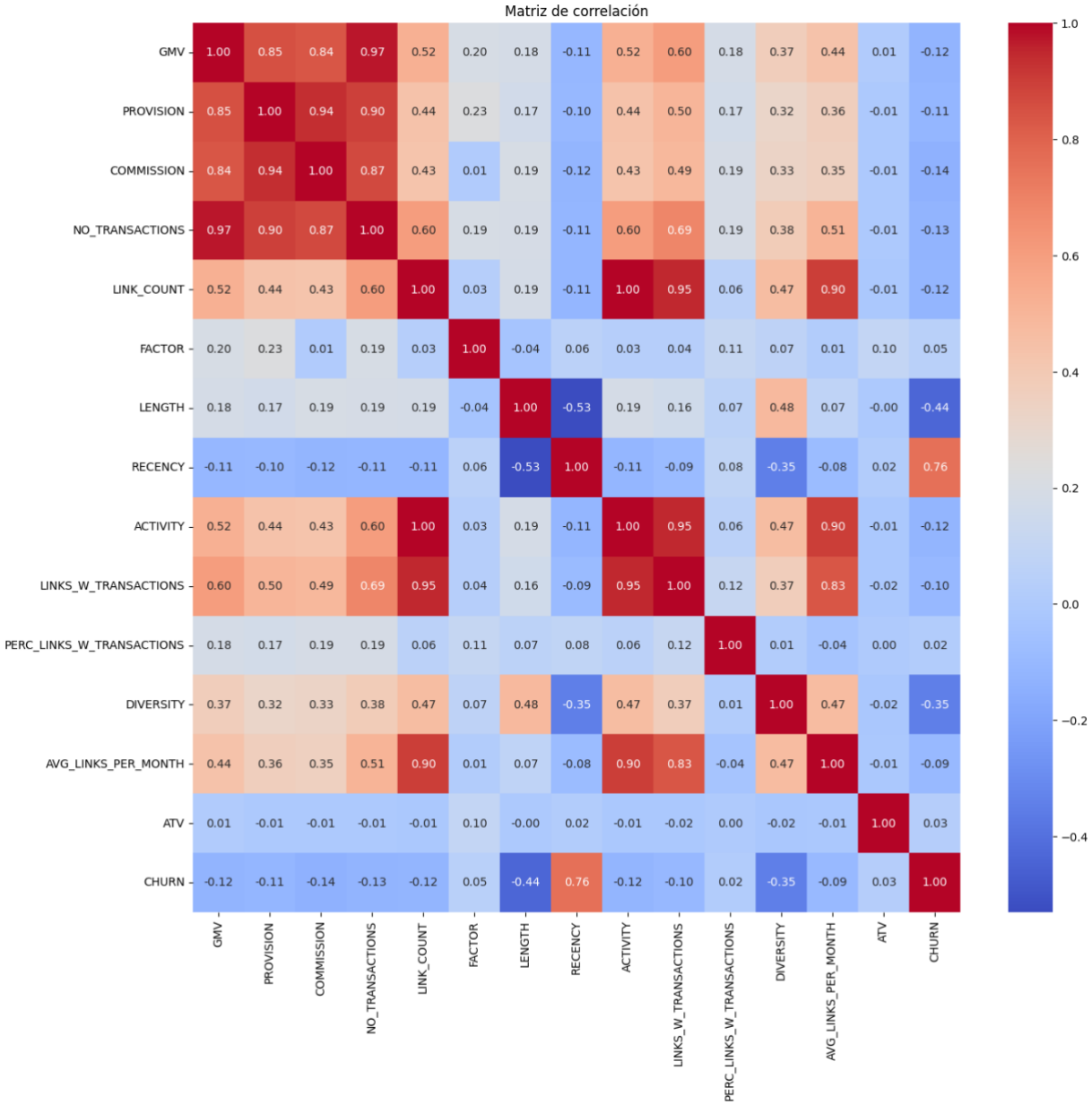
Cuanto mayor sea el número de enlaces con transacciones, mayor será el GMV generado. Por otra parte, si el último enlace ha sido publicado hace poco tiempo y el primero hace mucho, la diferencia entre el primer y el último enlace será mayor, por lo que la correlación en este caso es negativa.

Por último, se puede concluir que, a pesar de que el resultado de la matriz de correlación concuerda con la lógica, no se debe olvidar que, en la práctica, se dan relaciones entre variables que, a priori, podrían descartarse por no considerarse intuitivas. Por ello, para optimizar la exactitud de los modelos, es recomendable realizar una preselección de variables.

Como se ha mencionado en apartados anteriores, si las variables independientes presentan una correlación demasiado fuerte, proporcionarán información redundante que afectará al rendimiento del modelo.

Por el contrario, también se debe considerar que, cuando dos o más variables independientes tienen una baja correlación con la variable dependiente, estas no van a proporcionar ninguna información relevante en la fase de modelado. Si se incluyen en la fase de entrenamiento del modelo, este se verá obligado a aprender de forma incorrecta que las columnas tienen algún tipo de efecto sobre la variable dependiente.

Figura 12. *Matriz de correlación del conjunto de datos con las nuevas variables añadidas.*



Fuente: elaboración propia.

El resultado de no descartar variables con este tipo de correlaciones es un sobreajuste del modelo, de modo que las predicciones del conjunto de datos de entrenamiento serán

precisas, pero el rendimiento será bajo en la fase de validación (Misirlis & Vlachopoulou, 2021). En la fase de preprocesamiento, por lo tanto, se descartan aquellas variables que podrían causar un sobreajuste de los modelos.

3.2.1.2. Preprocesamiento de los datos

En la fase de análisis exploratorio, se han analizado los tipos de variables de las que se compone el conjunto de datos. Se ha observado que es necesario realizar una serie de cambios.

El USER_ID pasa a ser object, ya que se trata de un identificador que podría considerarse de tipo texto. REGISTER_DATE se cambia a formato de fecha (datetime64), LINKS_W_TRANSACTIONS pasa a ser de tipo numérico no decimal (int64), y CHURN pasa a ser booleano (bool), ya que los valores en esta variable solo pueden ser 0 (no churn) o 1 (churn). La Tabla 5 refleja los cambios aplicados a los tipos de variable.

Tabla 5. *Variables del conjunto de datos después de la transformación de tipos de variable.*

#	Column	Non-Null Count	Dtype
0	USER_ID	8296 non-null	object
1	REGISTER_DATE	8296 non-null	datetime64[ns]
2	GMV	8296 non-null	float64
3	PROVISION	8296 non-null	float64
4	COMMISSION	8296 non-null	float64
5	NO_TRANSACTIONS	8296 non-null	int64
6	LINK_COUNT	8296 non-null	int64
7	FACTOR	8296 non-null	float64
8	FIRST_LINK_DATE	8296 non-null	datetime64[ns]
9	LAST_LINK_DATE	8296 non-null	datetime64[ns]
10	LENGTH	8296 non-null	int64
11	RECENCY	8296 non-null	int64
12	ACTIVITY	8296 non-null	int64
13	LINKS_W_TRANSACTIONS	8296 non-null	int64
14	PERC_LINKS_W_TRANSACTIONS	8296 non-null	float64
15	DIVERSITY	8296 non-null	int64
16	AVG_LINKS_PER_MONTH	8296 non-null	float64
17	ATV	8296 non-null	float64
18	CHURN	8296 non-null	bool

dtypes: bool(1), datetime64[ns](3), float64(7), int64(7), object(1)

Fuente: elaboración propia.

Como se ha mencionado en el apartado anterior, es necesario descartar variables innecesarias para evitar el sobreajuste de los modelos. En este contexto, existen diferentes técnicas que se pueden aplicar. Una de ellas es el propio análisis de correlación en combinación con un

"conocimiento del dominio" o comprensión del negocio, que permite la eliminación de las variables innecesarias de forma manual. La experiencia en el negocio es fundamental para reconocer las variables pertinentes que permitirán predecir el abandono, como se ha demostrado en la fase de comprensión del negocio.

Otros métodos son la regularización L1 (Lasso), SelectKBest, el análisis de componentes principales (PCA) o RFE (Recursive Feature Selection for Feature Elimination) (Sharma, Patel & Shrivastava, 2023).

Este último se aplica en el presente trabajo debido a su uso habitual con algoritmos de aprendizaje supervisado (Sharma, Patel & Shrivastava, 2023). La RFE es una técnica recursiva que elimina sucesivamente una variable tras otra para comprobar cuál es el efecto en el rendimiento de los modelos con diferentes combinaciones de variables de modo que, finalmente, se pueda obtener la que mejor ajuste los modelos (Misirlis & Vlachopoulou, 2021).

En primer lugar, se eliminan las columnas no numéricas. Después, se determina la variable dependiente (CHURN) y se escalan las variables. Dado que en este trabajo se van a aplicar tanto el modelo de regresión logística como de árbol de decisión, se crean los modelos antes de aplicar RFE. Por último, se determinan cuántas variables se deben comprobar y se aplica la selección.

La literatura actual recomienda, en el caso de la regresión logística, incluir al menos 5 eventos positivos (Churn = 1) por cada variable independiente (EPV) (Vittinghoff & McCulloch, 2007). Con el tamaño de la muestra de stylink, donde se dan 3.945 eventos positivos, una selección de 10-15 variables independientes debería ser adecuada para no sobreajustar el modelo.

En el caso de los árboles de decisión, el número de variables no es tan relevante siempre que estén correlacionadas con la variable dependiente y no exacerben la complejidad del modelo. Esto debe mencionarse ya que, cuantas más variables se incluyan, más ramas y nodos se generarán (Hastie, Tibshirani, & Friedman, 2009) y menos interpretable será el modelo.

Como se muestra en la Tabla 6, tanto para la regresión logística como para el modelo de árbol de decisión, se seleccionan las variables: FACTOR, LENGTH, RECENCY, UNIQUE_DAYS, LINKS_W_TRANSACTIONS, y DIVERSITY.

Dado que el FACTOR captura implícitamente información sobre la provisión y la comisión, y el GMV está fuertemente relacionado con estas dos variables, las tres van a ser descartadas para el posterior modelado, de modo que se conserva solo FACTOR.

Por otra parte, LINKS_W_TRANSACTIONS y PERC_LINKS_TRANSACTIONS proporcionan la misma información, una en valores absolutos y la otra en valores relativos, de modo que se descarta la segunda.

Además, tanto RECENCY como CHURN miden el número de días transcurridos desde la última vez que un usuario creó un link. Para evitar el sobreajuste, se descarta RECENCY.

El resto de las variables que tienen una correlación al menos moderada según la matriz de correlación, sin llegar a la multicolinealidad con la variable dependiente, es decir, DIVERSITY y LENGTH, ya están incluidas en la selección.

Por lo tanto, para el modelo de regresión logística, las variables seleccionadas son: NO_TRANSACTIONS, FACTOR, LENGTH, ACTIVITY, LINKS_W_TRANSACTIONS, DIVERSITY. Para el árbol de decisión, las variables seleccionadas son: LINK_COUNT, FACTOR, LENGTH, ACTIVITY, LINKS_W_TRANSACTIONS, DIVERSITY, AVG_LINKS_PER_MONTH, ATV.

Tabla 6. *RFE para los modelos de Regresión Logística y Árbol de Decisión (en recuadro azul las variables comunes en los dos modelos).*

REGRESIÓN LOGÍSTICA	ÁRBOL DE DECISIÓN
-	LINK_COUNT
-	AVG_LINKS_PER_MONTH
NO_TRANSACTIONS	ATV
FACTOR	FACTOR
LENGTH	LENGTH
ACTIVITY	ACTIVITY
LINKS_W_TRANSACTIONS	LINKS_W_TRANSACTIONS
DIVERSITY	DIVERSITY

Fuente: elaboración propia.

El último paso en el preprocesamiento de datos consiste en la estandarización de las variables numéricas. La estandarización garantiza que todas las variables tengan una media de 0 y una varianza de 1, lo que estabiliza y acelera el entrenamiento de muchos algoritmos de aprendizaje automático.

Como se señaló anteriormente, algunas variables están sesgadas por la presencia de valores anómalos y, sin normalización, los modelos pueden requerir tasas de aprendizaje muy pequeñas, lo que lleva a tiempos de entrenamiento más largos. Al escalar los datos, mejoramos la convergencia y el rendimiento del modelo. (Misirlis y Vlachopoulou, 2021). Usamos StandardScaler de la biblioteca de Python scikit-learn.

Por último, el conjunto de datos limpio y preprocesado se utiliza en la siguiente fase del proyecto para el modelado de la regresión logística. Ya que los árboles de decisión no son tan sensibles a la distribución no normal de las variables, se utiliza el conjunto de datos no estandarizado para este modelo, lo que facilitará la lectura de los valores de las ramas del árbol.

3.3. Modelado propuesto

Una vez finalizada la limpieza y el preprocesamiento del conjunto de datos, la siguiente fase del proyecto según el modelo CRISP-DM consiste en el modelado (IBM, 2015). En este trabajo se proponen dos modelos de aprendizaje supervisado, la regresión logística y el árbol de decisión, por su uso habitual en el análisis churn, además de su interpretabilidad.

En primer lugar, los datos se dividen en subconjuntos de entrenamiento y prueba para evitar el sobreajuste, un problema común en el aprendizaje automático. El sobreajuste se produce cuando un modelo se ajusta demasiado a los datos de entrenamiento, de forma que no sabe generalizar a datos nuevos que no ha visto antes.

El conjunto de entrenamiento se usa para entrenar el modelo, mientras que el conjunto de prueba, que contiene nuevas observaciones, se usa para validarlo y evaluar su rendimiento. Si el modelo se comporta bien en el conjunto de prueba, puede considerarse validado (Misirlis y Vlachopoulou, 2021). Una división de datos común es 80/20, de modo que el 80 % de los datos forman el conjunto de entrenamiento y el 20 % el conjunto de prueba.

La principal diferencia entre la regresión logística y los árboles de decisión es que la regresión logística encuentra la mejor línea recta para separar los datos, mientras que los árboles de

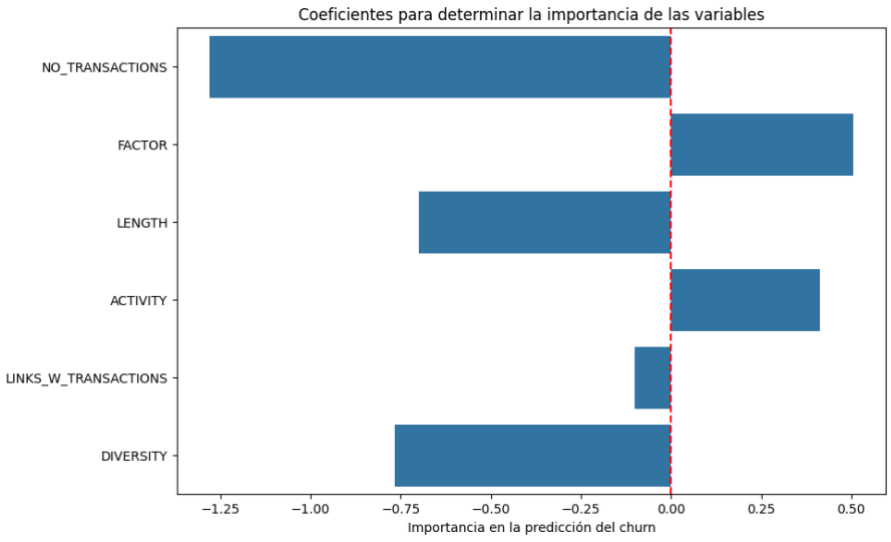
decisión dividen los datos en grupos más pequeños en función de la probabilidad de que ocurra un evento.

3.3.1. Regresión logística

La regresión logística es un tipo de análisis de regresión que se utiliza cuando la variable de salida es binaria. Se basa en un enfoque matemático que analiza el efecto de un conjunto de variables independientes sobre una variable dependiente. La predicción se realiza mediante la formulación de un conjunto de ecuaciones que conectan los valores de entrada (es decir, los factores que afectan a la pérdida de clientes) con el campo de salida (probabilidad de abandono). Se debe tener en cuenta que, en este tipo de modelos, la multicolinealidad puede llevar a conclusiones incorrectas sobre las relaciones entre variables, ya que otorga una magnitud errónea al coeficiente de regresión (Hassouna, Tarhini, Elyas, & Abou Trab, 2017).

Para realizar el modelado, primero se descargan las librerías y los paquetes necesarios, es decir, pandas, sklearn.model_selection y sklearn.linear_model. Se definen las variables independientes o predictoras (NO_TRANSACTIONS, FACTOR, LENGTH, ACTIVITY, LINKS_W_TRANSACTIONS, DIVERSITY) y la variable dependiente u objetivo (CHURN). Después, se crea el modelo de regresión logística y se entrena con el subconjunto de datos de entrenamiento. A continuación, se valida realizando predicciones con el conjunto de prueba. Como paso previo al análisis de la bondad del modelo, se representan gráficamente los coeficientes de las variables para comprobar el peso que tienen en la predicción del abandono.

Figura 13. *Coeficientes del modelo de regresión logística.*



Fuente: elaboración propia.

La Figura 13 se interpreta de manera que, en el caso de valores negativos, si disminuye el número de la variable predictora en una unidad, aumentará la posibilidad de churn por el valor las unidades del coeficiente (Kuhn & Johnson, 2013).

Por lo tanto, si disminuye el número de transacciones, de links con transacciones, la diversidad de los enlaces y el número de días entre el último y el primer link, aumentará la posibilidad de churn en el número de unidades de los coeficientes respectivos. Por otro lado, cuanto mayor sea el factor y más aumente el número de días únicos en los que los usuarios crean enlaces, más disminuye la posibilidad de abandono.

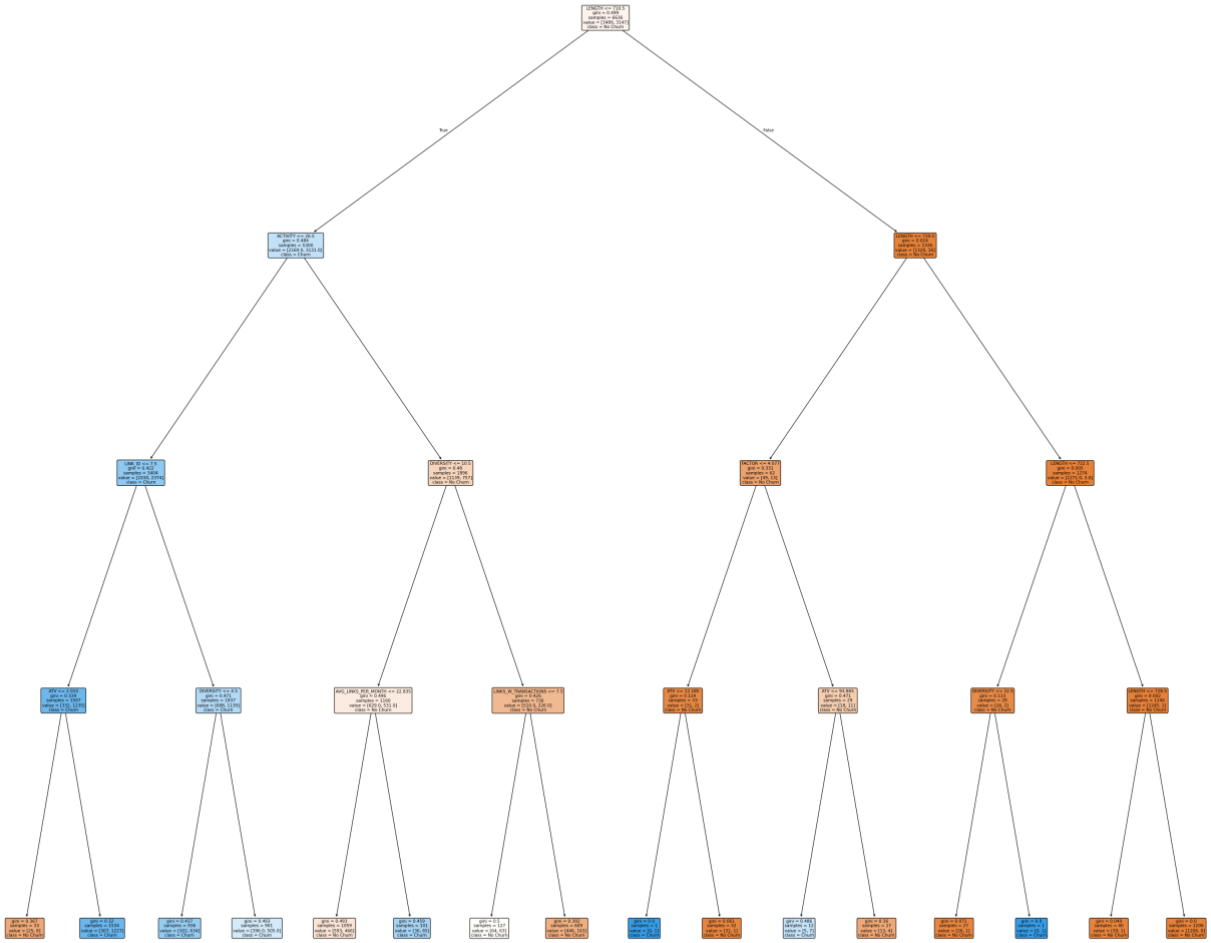
3.3.2. Árbol de decisión

El árbol de decisión es un modelo que genera una estructura en forma de árbol para representar un conjunto de decisiones. Devuelve las probabilidades de pertenecer a una clase (churn o no churn). Se compone de nodos internos, ramas y nodos hoja. Los nodos internos representan una sola variable y constituyen un punto de prueba. Las ramas representan el resultado de la prueba y forman líneas que conducen a los nodos hoja, que representan las etiquetas de las clases. Se trata de un modelo flexible e interpretable, ya que permite el uso de datos categóricos y continuos (Sabbeh, 2018).

Para realizar el modelado, se descargan en primer lugar las librerías necesarias, sklearn.tree, sklearn.model_selection y matplotlib.pyplot. Se definen de nuevo las variables predictoras (LINK_COUNT, FACTOR, LENGTH, ACTIVITY, LINKS_W_TRANSACTIONS, DIVERSITY, AVG_LINKS_PER_MONTH, ATV) y la variable objetivo (CHURN).

A continuación, se crea el modelo de árbol de decisión y se define la profundidad del árbol. Esta definición se puede realizar de forma aleatoria o con una validación cruzada, la cual se aplica al presente modelo mediante la librería sklearn.model_selection y el paquete GridSearchCV. GridSearchCV implementa un método de ajuste y un método de puntuación, entre otros, para encontrar los mejores parámetros que se ajusten a los respectivos modelos (Scikit-learn developers, s.f.). Prueba de forma automática distintas profundidades y selecciona la que mejor funciona. En este caso, la profundidad con los mejores resultados es 4. Finalmente, se ajusta el modelo con los datos de entrenamiento y los parámetros seleccionados y se aplica al conjunto de prueba. Por último, se visualizan los resultados, como se aprecia en la Figura 14.

Figura 14. *Resultado del modelo de árbol de decisión.*



Fuente: elaboración propia.

Dada la magnitud del gráfico, se incluye el detalle del modelo en el Anexo B. Los nodos y las ramas se interpretan de la forma que se expone a continuación.

El primero nodo, LENGTH <= 710.5, gini = 0.499, samples = 6636, value = [3489, 3147], class = no churn, indica que el modelo divide a los influencers según la variable LENGTH, en función de si el número de días transcurrido entre su primer y último link es mayor o menor de 710,5 días (se redondea el valor). El valor “Gini” indica que los influencers churn y no churn están mezclados casi al 50% en el nodo. “Samples” indica que hay 6.636 influencers incluidos en el nodo. De estos, “value” indica que 3.489 no son churn y 3.147 son churn. Por último, “class”

proporciona la información adicional de que el nodo clasifica según los usuarios no churn, ya que este es el grupo mayoritario.

La interpretación de este primer nodo implica, en definitiva, que cuando el tiempo transcurrido entre el primer y último link de un usuario es menor o igual a 710 días, hay una posibilidad mayor de que estos no abandonen. La primera rama divide a los usuarios entre aquellos con una LENGTH menor o igual a 710,5, por una parte, y mayor a 710,5 por otra.

Cabe destacar que la mayor parte de los demás nodos parten de un Gini relativamente cercano a 50%, con lo que las muestras están mezcladas de forma homogénea entre usuarios churn y no churn.

El resto de los nodos se interpreta de la misma forma que el primero. Por ejemplo, si se sigue la primera rama de la izquierda, el modelo clasifica con la variable ACTIVITY <= 26.5, gini = 0.484, samples = 5300, value = [2169, 3131], class = churn. Se clasifica a los usuarios churn en función de si el número de días únicos en el que han publicado enlaces es menor o igual a 26,5. De los 5.300 usuarios que componen el nodo, 2.169 pertenecen a la clase no churn y 3.131 a la clase churn, y esta constituye en este caso la mayoría.

En conclusión, si los usuarios han publicado su último link con 710 o más días de diferencia con el primero, y han estado activos 26 días o menos, tienen una mayor probabilidad de churn.

Si se sigue analizando el árbol en todos sus niveles de profundidad, se pueden realizar diversas afirmaciones.

Este primer grupo de usuarios procedente del primer nodo y la primera rama que conduce al segundo nodo por la izquierda, también serán churn si su número de enlaces total es 7 o menos, y su ATV es igual o menor a 2€.

El segundo grupo de usuarios con una LENGTH inferior o igual a 710 días y una ACTIVITY superior a 26 días, tienden a no abandonar cuando la diversidad de las marcas que recomiendan es menor o igual a 10 y su media de enlaces al mes es de 22 o menos.

Si se analiza a partir de la primera rama por la derecha, el tercer grupo de usuarios con una LENGTH superior a 710 días, pero menor o igual a 719 días, con un FACTOR menor o igual a 4 y un ATV menor o igual a 12€ tienden a no abandonar.

Los últimos nodos de la parte inferior son los nodos hojas, que indican el final del árbol, y no clasifican a los usuarios restantes en función de ninguna variable.

En la Tabla 7 se presentan de forma resumida las variables que han clasificado a una mayoría de usuarios churn, así como sus características generales según los resultados del modelo.

Tabla 7. *Características de los usuarios perdidos según el modelo de árbol de decisión.*

USUARIOS CHURN	
ACTIVITY	Han publicado links en 26 días únicos o menos
LINK_COUNT	Han creado 7 links o menos entre enero de 2023 y abril de 2025
DIVERSITY	Han creado enlaces de 4 o menos marcas distintas
ATV	Sus links han generado un ATV igual o inferior a 2€

Fuente: elaboración propia.

En el apartado 4 se evalúan los modelos y se analiza en profundidad su rendimiento mediante diversas métricas de validación.

4. Construcción, prueba, implementación y despliegue

4.1. Medidas de bondad del modelado

En el apartado anterior se crean los modelos, se entrenan y se realizan predicciones con el conjunto de datos de entrenamiento y de prueba, respectivamente. A continuación, se analiza su calidad a través de diferentes métricas y se elige el modelo con un mejor rendimiento, lo que corresponde a la fase 5 del proyecto según el modelo CRISP-DM (IBM, 2015).

4.1.1. Métricas de validación del modelo de regresión logística

Para evaluar el modelo de regresión logística, se eligen cinco métricas que están incluidas en la librería sklearn.metrics: la exactitud, la matriz de confusión, el informe de clasificación, la curva ROC y la curva de aprendizaje. La Figura 15 ilustra estas métricas.

Todas estas métricas se basan en tuplas de positivos y negativos. Dada nuestra variable dicotómica, CHURN sí o no, los sí se refieren a los valores positivos y los no a los negativos. Los modelos pueden predecir los valores, tanto positivos como negativos, de forma correcta o errónea.

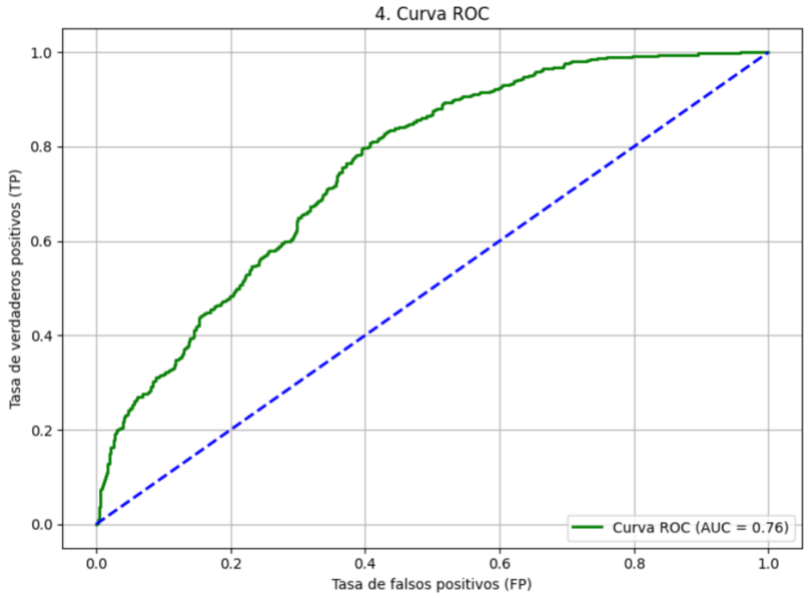
De esta combinación se obtienen cuatro métricas: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN) (Han, Kamber, & Pei, 2011).

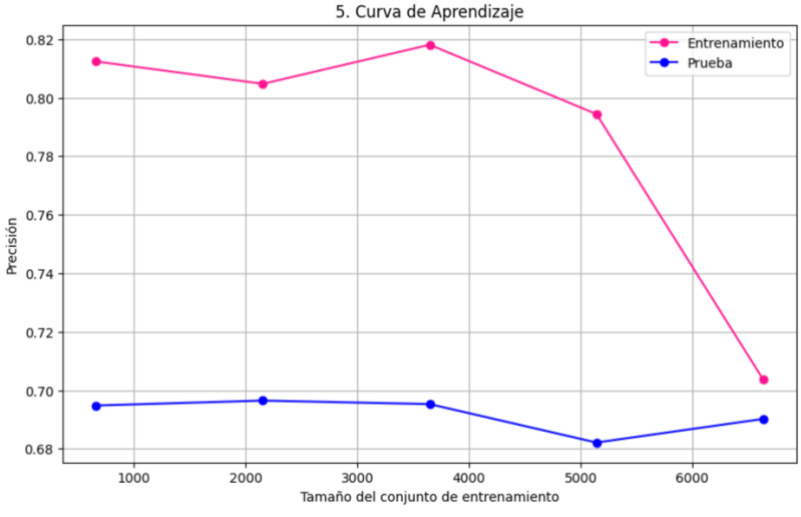
Figura 15. *Métricas de validación del modelo de regresión logística.*

```
1. Exactitud del modelo: 0.6746987951807228
2. Matriz de confusión:
 [[573 289]
 [251 547]]
3. Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.70      0.66      0.68       862
           1       0.65      0.69      0.67       798

    accuracy                           0.67      1660
   macro avg       0.67      0.68      0.67      1660
weighted avg       0.68      0.67      0.67      1660
```





Fuente: elaboración propia.

La exactitud o accuracy se calcula dividiendo la tasa de verdaderos positivos y negativos entre todos los valores predichos (Powers, 2011), es decir, cuántos influencers el modelo ha predicho de forma correcta. Cuanto más cerca de 100% esté la métrica, mayor será la capacidad de predicción del modelo. El resultado obtenido es 0,6746 (67,5%), lo que indica que se trata de un modelo con cierta capacidad predictiva, pero con posibilidad de optimización.

Todas las combinaciones mencionadas se encuentran resumidas en la matriz de confusión, que es una herramienta que se utiliza para analizar la bondad del modelo para reconocer las tuplas de las diferentes clases (Han, Kamber, & Pei, 2011). Según los resultados obtenidos, el modelo predijo de forma correcta que 573 influencers no abandonan (TN) y 547 sí abandonan (TP). Por otra parte, erró al predecir 289 falsos abandonos (FP) y 251 falsos no abandonos (FN).

El reporte de clasificación proporciona información acerca de la precisión, el Recall y la F1-Score. Precisión hace referencia al número de verdaderos positivos divididos entre la suma de verdaderos positivos más falsos positivos. “Recall” o sensibilidad es la tasa de verdaderos positivos, es decir, los verdaderos positivos divididos entre la suma de verdaderos positivos y falsos negativos. Por último, F1- Score es la media armónica de la precisión y el Recall y se obtiene multiplicando la precisión por el Recall y por 2, y dividiendo entre la precisión más el Recall (Han, Kamber, & Pei, 2011).

Los resultados obtenidos indican, con un 70% versus un 65%, que el modelo es algo más preciso a la hora de predecir el no churn, pero el Recall ligeramente más alto en churn, 69% versus 66%, indica que también identifica de forma adecuada a los churn. Por último, la media armónica de las dos métricas, 67% (churn) – 68% (no churn), indica que ambas están equilibradas.

La curva ROC permite visualizar el equilibrio entre la sensibilidad y la tasa de falsos positivos a lo largo de diferentes umbrales del conjunto de prueba. El AUC, o área bajo la curva, se calcula dividiendo los verdaderos entre los falsos positivos, y es también una medida utilizada para la validación (Han, Kamber, & Pei, 2011). El AUC de 0,76 obtenido indica una buena capacidad discriminativa entre los usuarios churn y no churn.

En la Tabla 8 se presentan las fórmulas de las métricas utilizadas para evaluar el modelo de regresión logística.

Tabla 8. *Métricas de validación de los modelos.*

MÉTRICA DE VALIDACIÓN	FÓRMULA
Exactitud	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$
Precisión	$\text{Precision} = \frac{TP}{TP + FP}$
Recall	$\text{Recall} = \frac{TP}{TP + FN}$
F1-Score	$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
AUC	It plots TP rate against FP rate

Fuente: elaboración propia.

Por último, las curvas de aprendizaje son gráficos del rendimiento del modelo en el conjunto de entrenamiento y en el conjunto de prueba. Para generarlos, se entrena el modelo varias veces con subconjuntos de distintos tamaños que se extraen del conjunto de entrenamiento (Geron, 2019). Se utilizan para evaluar si los modelos se ajustan bien a los datos.

La curva de aprendizaje obtenida para el modelo de regresión logística con los datos de entrenamiento muestra una exactitud del 82% con un tamaño de aproximadamente 3.900 influencers y cae drásticamente hasta el 70%, como se puede observar en la Figura 15, al incrementar el conjunto a más de 6.000 usuarios. Esto supone que el modelo no está aprendiendo a medida que se enriquece con más datos, lo que podría estar causado por una falta de linealidad. Por otra parte, la curva del conjunto de prueba se mantiene estable alrededor del 69% hasta los 3.900 influencers, cae al 68% con algo más de 5.000 influencers en la muestra, y vuelve a subir ligeramente con una muestra de más de 6.000. Esto implica que, si bien el modelo no está sobreajustado, tampoco mejora su rendimiento al incrementar el tamaño de la muestra y, además, no es capaz de identificar patrones más complejos en la dinámica del churn.

4.1.2. Métricas de validación del modelo de árbol de decisión

Los resultados obtenidos al realizar la regresión logística dan indicios de la necesidad de implementar un modelo en el que la linealidad no sea un requisito para un buen rendimiento. Para evaluar el modelo de árbol de decisión, se eligen de nuevo las métricas de exactitud, la matriz de confusión, el informe de clasificación y la curva de aprendizaje.

En primer lugar, la exactitud de 0.7139 indica que el modelo clasificó adecuadamente a un 71,4% de los influencers que no abandonan. Esta métrica mejora la alcanzada por el modelo de regresión logística.

Por otra parte, la matriz de confusión indica que el modelo predijo correctamente a 578 influencers que no abandonan y a 607 que sí abandonan. En cambio, se equivocó al clasificar como churn a 284 que en realidad no abandonaron y en no churn a 191 que sí abandonaron. Estos resultados mejoran también ligeramente los obtenidos en la anterior matriz de confusión.

El informe de clasificación muestra una precisión del 75% para no churn y 68% para churn. La métrica Recall señala que el modelo identificó correctamente a un 67% de los influencers que no abandonaron y a un 76% de los que sí abandonaron, es decir, que el modelo es capaz de identificar ambas clases de manera adecuada, pero funciona mejor a la hora de identificar a los usuarios con potencial de churn. La F1-Score de 71-72% también indica un equilibrio entre ambas métricas, que son de nuevo superiores a las del modelo de regresión logística.

Por último, la curva de aprendizaje para el modelo con datos de entrenamiento comienza con una precisión de algo más de 0.76, desciende a 0.74 a partir de un tamaño de muestra de 2.000 influencers y se mantiene estable. Esto implica que el modelo no necesita de un gran conjunto de datos para aprender los patrones principales del comportamiento de los influencers churn. La curva del conjunto de datos de prueba, en cambio, no es tan constante. Comienza con una precisión de aproximadamente 0.66 y se eleva a 0.68 al alcanzar la muestra con 2.000 usuarios. Después, desciende significativamente por debajo de 0.66 al aumentar el tamaño a algo menos de 4.000 usuarios. La precisión mejora por encima de 0.68 al incorporar una muestra con algo más de 5.000 influencers y vuelve a oscilar por debajo de 0.68 cuando el tamaño es mayor de 6.000. Esta fluctuación implica que el modelo no es tan preciso a la hora de clasificar nuevos datos y, a su vez, que es sensible a la varianza de las variables

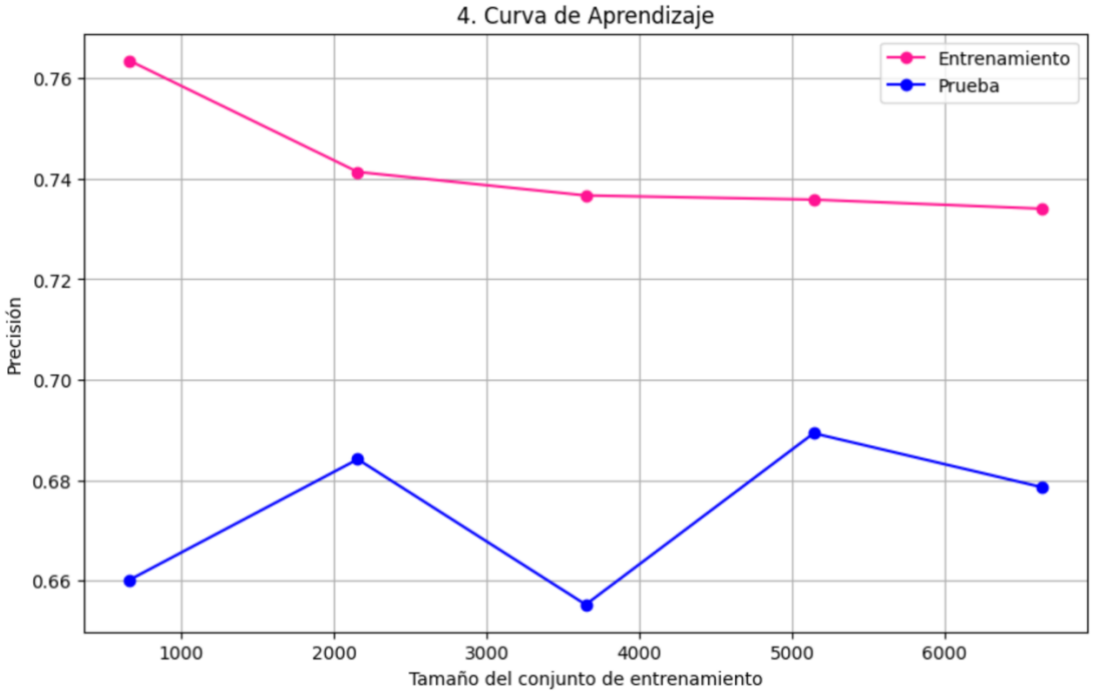
independientes. En el apartado de limitaciones se sugieren métodos para mejorar estas métricas. La Figura 16 resume los resultados obtenidos al evaluar el modelo de árbol de decisión.

Figura 16. *Métricas de validación del modelo de árbol de decisión.*

```
1. Exactitud del modelo: 0.7139
2. Matriz de confusión:
[[578 284]
 [191 607]]
3. Reporte de clasificación:
              precision    recall  f1-score   support

       False       0.75      0.67      0.71       862
        True       0.68      0.76      0.72       798

    accuracy                           0.71      1660
   macro avg       0.72      0.72      0.71      1660
weighted avg       0.72      0.71      0.71      1660
```



Fuente: elaboración propia.

En conclusión, después de analizar ambos modelos, y tras comprobar que todas las métricas de validación del árbol de decisión superan a las del modelo de regresión logística, se descarta este último, y se procede a la implementación del modelo de árbol en el siguiente apartado.

4.2. Ejemplo de aplicación

Este apartado corresponde a la fase 6 del modelo CRISP-DM (IBM, 2015). Una vez elegido el modelo final, se extraen los USER_ID que el modelo ha identificado como usuarios churn y se calcula su probabilidad de abandono, como se representa en la Figura 17, con el comando predict_proba. Finalmente, como se puede apreciar en la Figura 18, se representa su distribución con un histograma, con el objetivo de averiguar en qué zona de riesgo se encuentran.

Se observa que, de los 891 usuarios predichos como churn en el conjunto de datos de prueba, un 43% se encuentra en un muy elevado riesgo de abandono (distribución alrededor de 0.8), el 24% en un alto riesgo (alrededor de 0.7), y el 33% en un nivel de medio-alto riesgo (alrededor de 0.6).

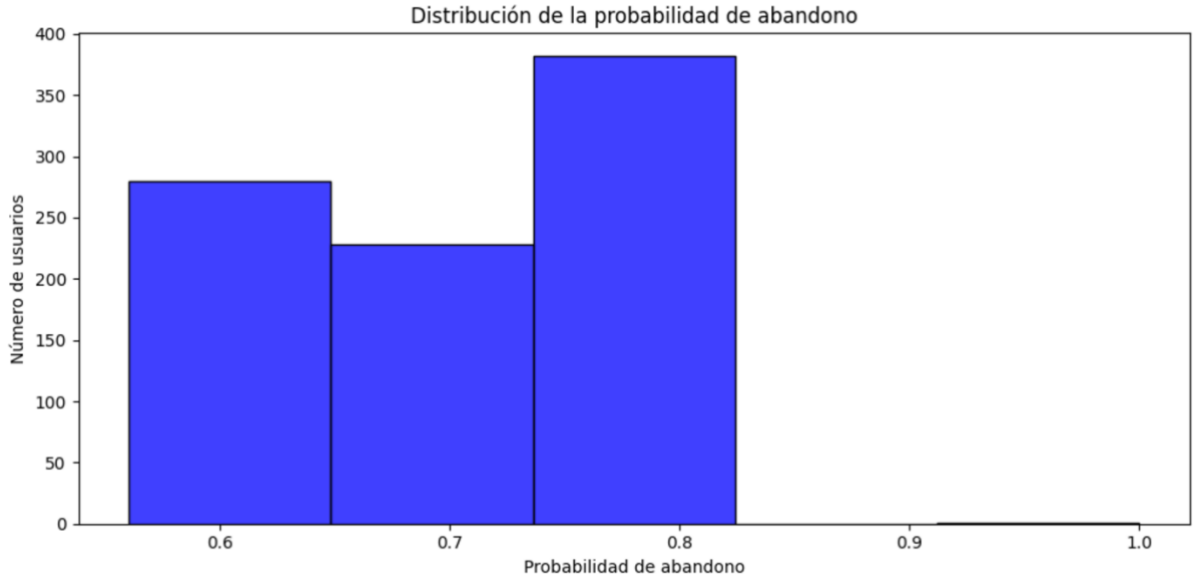
Será necesario, por tanto, implementar una estrategia de retención que priorice a los usuarios que aporten a la empresa valor monetario y que se encuentren en un umbral de riesgo de abandono más elevado. Además, se recomienda a los principales stakeholders del presente trabajo, la vicepresidenta de finanzas y los responsables del mercado alemán, tener en cuenta las variables que el modelo ha identificado como más relevantes para el diagnóstico churn.

Figura 17. *Lista de USER_ID clasificados como Churn y su probabilidad de abandono.*

USER_ID	CHURN_PROBABILITY
157196	56,05%
279944	79,99%
94745	56,05%
281790	67,74%
125323	56,05%
176365	79,99%
138146	56,05%
126259	58,33%
84583	79,99%
53601	56,05%
4646	67,74%
266749	56,05%
33869	67,74%
142878	67,74%
172615	56,05%
164184	56,05%
147352	67,74%
188936	79,99%
91077	56,05%
198925	79,99%
170951	79,99%

Fuente: elaboración propia.

Figura 18. *Histograma de la distribución de los usuarios churn en función de su probabilidad de abandono.*



Fuente: elaboración propia.

Se deberá observar especialmente el número de enlaces creados por los influencers, tanto a nivel absoluto como en días únicos, el valor monetario medio generado por cada transacción, así como la diversidad de marcas que promocionan a través de sus links. Si todos estos factores se encuentran en el umbral crítico observado por el modelo de árbol y, además, el modelo ha clasificado a dichos usuarios como churners potenciales, stylink deberá considerar si aportan un valor suficiente para que sea rentable invertir recursos monetarios en mantenerlos activos.

Finalmente, algunas medidas que los responsables del mercado alemán pueden implementar para lograr este objetivo son, entre otros, incrementar la compensación por clic que el influencer percibe, ofrecerle la participación en colaboraciones con sus marcas favoritas, códigos de descuento para sus seguidores, o invitarlos a los eventos para creadores de contenido que stylink organiza a nivel internacional.

5. Cronograma del proyecto

Los swimlanes se utilizan para dividir y organizar las tareas de un proyecto en forma de diagrama. Se componen de “piscinas”, que representan los procesos que lo conforman, y de “pistas”, que definen los roles y las tareas de los participantes del proyecto dentro de los respectivos departamentos implicados (White & Miers, 2008).

5.1. Swimlane de la gestión del proyecto

El swimlane de la gestión del proyecto engloba la coordinación de todas sus etapas por parte del equipo de datos junto con los principales stakeholders. Primero, se define el caso de estudio, la metodología a seguir, el conjunto de datos a analizar, las fases, así como el cronograma de proyecto, que se llevará a cabo entre marzo y julio de 2025. La científica de datos del equipo y la vicepresidenta de finanzas se encargan de revisar el cumplimiento de plazos a lo largo de todas las etapas.

5.2. Swimlane de datos y bases de datos

En este swimlane se incluyen todas las actividades relacionadas con la recopilación de los datos históricos de los influencers de stylink. Se accede al almacén Snowflake y se extrae el conjunto de datos que se utiliza para el posterior análisis y modelado. Este incluye información transaccional relativa al comportamiento de los influencers con sus enlaces de afiliación y las ganancias generadas por estos para las marcas asociadas. El equipo de datos se encarga de la extracción y no se requiere el apoyo de ningún otro equipo.

5.3. Swimlane de la integración de datos

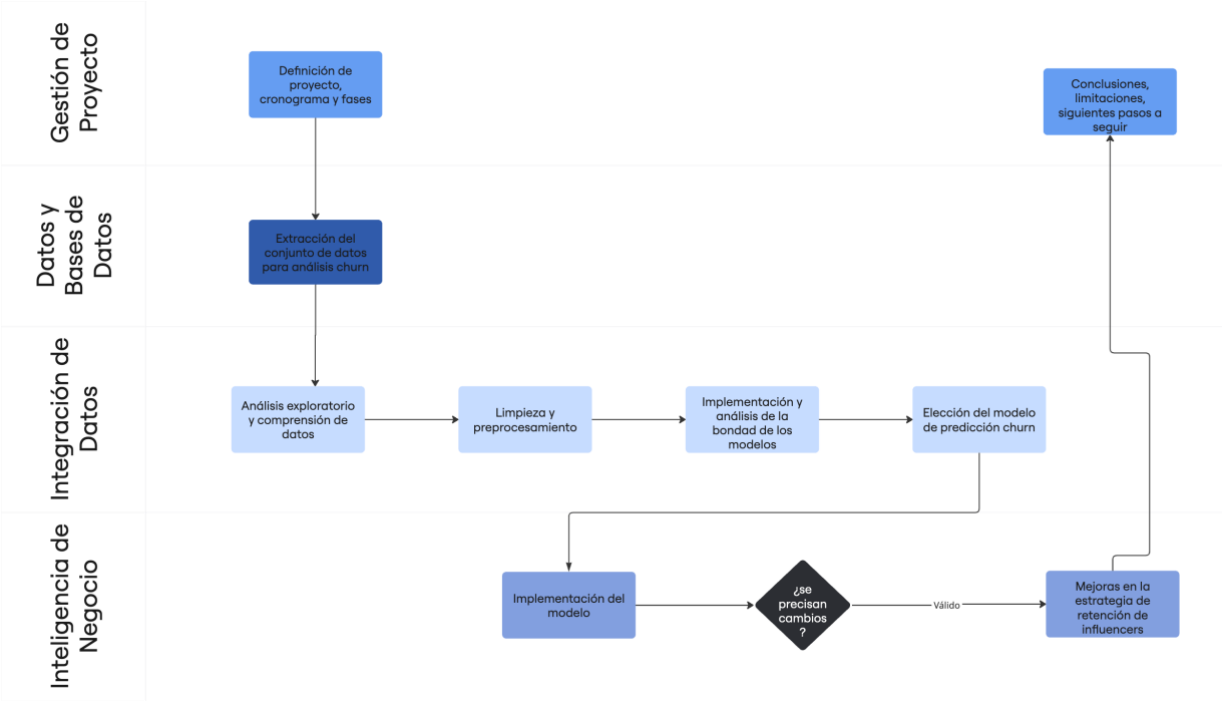
La integración comprende el análisis exploratorio de las variables incluidas en el conjunto de datos extraído, la comprensión de estos, así como todo tipo de tareas relacionadas con la limpieza y el preprocesamiento que los modelos requieren para obtener un buen rendimiento. Finalmente, en esta fase, se aplican los modelos y se selecciona la mejor variante. De nuevo, el equipo de datos es el único implicado en esta fase.

5.4. Swimlane de Inteligencia de Negocio

Este swimlane conlleva la implementación de los resultados. Se predice la probabilidad churn de los influencers, y se proponen recomendaciones para implementar una estrategia de retención basada en los hallazgos obtenidos. Finalmente, se exponen las conclusiones y limitaciones a los stakeholders, y se definen los pasos a seguir en los siguientes análisis churn.

En la Figura 19, se presenta de forma gráfica el swimlane de proyecto del presente trabajo.

Figura 19. *Swimlane de proyecto de análisis churn de los influencers de stylink.*



Fuente: elaboración propia.

6. Conclusiones

El presente trabajo se concibe con el objetivo de analizar el comportamiento de los influencers de stylink en el mercado alemán, e identificar patrones para predecir su abandono. Para ello, se accede a datos reales de la empresa de los últimos años y se extrae información relativa al historial transaccional de los usuarios. En primer lugar, se realiza un estudio del contexto actual en el marketing de afiliación y de influencers para comprender el origen y el alcance de la problemática experimentada por las empresas en este sector.

A continuación, se explora la literatura existente con respecto a la predicción churn. Se observa que el uso de algoritmos de aprendizaje en conjunto proporciona en general un mejor rendimiento; no obstante, para el caso de stylink, se seleccionan la regresión logística y los árboles de decisión por su uso habitual en la industria, su flexibilidad y su interpretabilidad.

Como paso previo al modelado, se realiza un análisis exploratorio del conjunto de datos con el objetivo de extraer los factores capaces de explicar los patrones de abandono. Se obtienen variables que recogen información acerca de la frecuencia, el valor monetario y la diversidad de los enlaces y las transacciones de los influencers. Se analizan las correlaciones para evitar problemas de multicolinealidad y se procede a la limpieza y el preprocesado de datos.

El modelo de regresión logística obtiene una exactitud del 67,5% y un AUC de 0,76, lo que supone un sólido punto de partida para la discriminación entre usuarios churn y no churn. El resto de las métricas usadas para la evaluación del modelo confirman su plausibilidad. Sin embargo, la curva de aprendizaje señala indicios de problemas en el modelo a la hora de identificar patrones complejos en el comportamiento de los usuarios churn, posiblemente debido a las relaciones no lineales que se dan entre las variables.

El modelo de árbol de decisión, por su parte, supera en rendimiento al anterior en todas las métricas. Obtiene una exactitud del 71,4% y una F1-Score del 72%, y permite identificar de forma más precisa las variables que influyen en el abandono de los influencers. Este resultado respalda los hallazgos expuestos en la literatura con respecto a la bondad de los árboles de decisión para este tipo de análisis. Sería conveniente, en previsión a futuros trabajos, comparar este con modelos de árbol de decisión en conjunto, potenciados por gradiente, o incluso bosques aleatorios para seguir mejorando el rendimiento de este.

Por otra parte, gracias a la interpretabilidad de este modelo, se descubre que factores como una baja cantidad de enlaces, tanto en valores totales como mensuales, un reducido número de marcas promocionadas y un valor transaccional promedio bajo por link son características comunes de los usuarios churn. La aplicación de modelos más complejos podría enriquecer también estos resultados al identificar nuevos patrones en las variables.

Como paso final, se extrae la información de los usuarios que el modelo ha clasificado como churn en el conjunto de datos de prueba y se predice su probabilidad churn para poder elaborar una serie de recomendaciones a implementar en la estrategia de retención de stylink.

Se recomienda a la empresa dividir a los usuarios en función de su riesgo de abandono. Para aquellos con un riesgo más elevado, se propone llevar a cabo medidas de retención más personalizadas y con un alcance más inmediato como, por ejemplo, un incremento de la comisión por clic para incentivar una actividad constante.

Los que se encuentran en un umbral de riesgo medio deben ser tratados con cautela, puesto que una experiencia de cliente negativa podría suponer el abandono inmediato; por ello, se recomiendan medidas con un alcance a medio plazo, por ejemplo, la participación de los influencers en colaboraciones exclusivas con marcas asociadas, o una invitación a los eventos para creadores de contenido organizados por stylink.

Por último, para los influencers con un riesgo bajo de abandono, se recomienda no invertir recursos monetarios de forma inmediata, sino observar las variables que los modelos han identificado como decisivas para la predicción churn, y mantener un servicio de atención al cliente más personalizado, a fin de identificar de forma preventiva señales de insatisfacción y poder subsanarlas antes de que el riesgo aumente.

En conclusión, este estudio demuestra que es posible predecir con una exactitud elevada el riesgo de abandono de los usuarios de stylink a través del uso de datos históricos y la implementación de modelos de aprendizaje supervisado.

Además, también se identifican las variables que más influyen en el incremento de la tasa churn, de modo que los respectivos responsables podrán implementar una estrategia de retención adaptada a los requisitos y necesidades de los influencers del mercado alemán.

7. Limitaciones y prospectiva

Es necesario señalar que este trabajo presenta una serie de limitaciones. En primer lugar, se toma como marco temporal un periodo, de enero de 2023 a abril de 2025, que no engloba todos los datos históricos de los influencers de stylink Alemania. Esto implica que los patrones de comportamiento analizados podrían no ser representativos para aquellos que se registraron antes del 2023.

Por otra parte, la elección de los modelos de regresión logística y árbol de decisión dado su habitual uso y su interpretabilidad no implica que estos sean también óptimos. Como se ha expuesto en el apartado 2, existen otros algoritmos más complejos con los que se podría obtener un mejor rendimiento.

Además, las métricas usadas para evaluar los modelos podrían haber mejorado con cambios en las variables que se incorporaron en el análisis o con un aumento del tamaño de la muestra y del número de iteraciones en la fase de validación.

En cuarto lugar, se exploran una serie de variables independientes que están estrechamente relacionadas con el rendimiento monetario de la empresa y el historial transaccional de los influencers. Sin embargo, otros factores potencialmente relevantes, como las características demográficas (por ejemplo, el género, la edad, el nivel educativo) o las métricas específicas

de los influencers (por ejemplo, el número de seguidores, el nicho de mercado, o la tasa de interacción) no se incluyen en este trabajo y podrían ofrecer un conocimiento valioso.

Por último, el presente trabajo está enfocado en un caso empresarial real y se enmarca en un nicho de mercado específico del marketing de influencers, de modo que podría ser difícil generalizar el modelo y aplicarlo a estudios empresariales en otros sectores.

Con vistas a futuros análisis, se podrían considerar otras constelaciones temporales que incluyan una muestra mayor de usuarios, además de realizar un análisis churn en otros mercados internacionales en los que stylink está experimentando un incremento en la tasa de abandono.

Además, para mejorar el rendimiento, se deberían explorar otros algoritmos más complejos de aprendizaje en conjunto, como los bosques aleatorios o los árboles de decisión potenciados por gradiente. Estos modelos han demostrado una precisión superior en problemas de predicción similares y podrían revelar nuevos matices en el comportamiento de los usuarios.

Por otra parte, sería necesario explorar otro tipo de variables independientes para comprobar el efecto que estas tienen sobre la variable dependiente. Podría existir una relación estrecha, por ejemplo, entre la tasa de abandono y el número de seguidores de los influencers, o incluso su rango de edad.

Por último, si se deseara adaptar el presente trabajo a otro caso de estudio churn, se debería prácticamente realizar un nuevo análisis exploratorio y contextual para ajustar los modelos a las singularidades de los diferentes modelos de negocio en la industria del marketing de influencers y de afiliación.

Referencias bibliográficas

Amazon Web Services. (s. f.). *Snowflake Partner Solutions for Financial Services* [Gráfico]. AWS. https://aws.amazon.com/de/financial-services/partner-solutions/snowflake/

Campbell, C., & Farrell, J. R. (2020). More than meets the eye: The functional components underlying influencer marketing. *Business Horizons, 63*(4), 469–479.

Dajah, S. (2020). Marketing through social media influencers. *International Journal of Business and Social Science, 11*(9), 71. https://doi.org/10.30845/ijbss.v11n9p9

Fader, P. (2012). *Customer centricity: Focus on the right customers for strategic advantage*. Wharton Digital Press.

Fader, P., Hardie, B., & Ross, M. (2022). *The customer-base audit: The first step on the journey to customer centricity*. Wharton School Press.

Fader, P., & Toms, S. (2018). *The customer centricity playbook: Implement a winning strategy driven by customer lifetime value*. Wharton School Press.

Geron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

Gürel, E., & Tat, M. (2017). SWOT analysis: A theoretical review. *Journal of International Social Research, 10*(51), 994–1006. https://doi.org/10.17719/jisr.2017.1832

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hassouna, M., Tarhini, A., Elyas, T., & Abou Trab, M. S. (2015). Customer churn in mobile markets: A comparison of techniques. *International Business Research, 8*(6), 224. https://doi.org/10.5539/ibr.v8n6p224

Hassouna, M., Tarhini, A., Elyas, T., & Abou Trab, M. S. (2017). Customer churn in mobile markets: A comparison of techniques. *Journal of Business Research, 72*, 205–211. https://doi.org/10.1016/j.jbusres.2017.02.042

Höppner, S., Stripling, E., Baesens, B., Vanden Broucke, S., & Verdonck, T. (2017). *Profit-driven decision trees for churn prediction*. KU Leuven & University of Southampton.

Jaramillo-Chuqui, I. F., & Villarroel-Molina, R. (2023). *Elementos básicos de análisis inteligente de datos*. Editorial Grupo AEA. https://doi.org/10.55813/egaea.l.2022.65

Kaushik, A. (2009). *Web analytics 2.0: The art of online accountability & science of customer centricity*. Wiley.

Kemp, S. (2021). *Digital 2021: Global overview report*. We Are Social and Hootsuite. https://wearesocial.com/us/blog/2021/01/digital-2021-the-latest-insights-into-the-state-of-digital/

Kim, S., & Lee, H. (2021). Customer churn prediction in influencer commerce: An application of decision trees. *Procedia Computer Science, 199*, 1332–1339. https://doi.org/10.1016/j.procs.2022.01.169

Kim, S., & Lee, H. (2021). Customer churn prediction in influencer commerce: An application of decision trees. *The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)*.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

Mishra, A., & Reddy, U. S. (2017). A comparative study of customer churn prediction in telecom industry using ensemble-based classifiers. En *Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)*. IEEE Xplore.

Misirlis, N., & Vlachopoulou, M. (2021). *Data science for marketing analytics* (2nd ed.). Packt Publishing.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Rachid, A. D., Abdellah, A., Belaid, B., & Rachid, L. (2018). Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context. *International Journal of Electrical and Computer Engineering, 8*(4), 2367.

Raeisi, S., & Sajedi, H. (2020). E-commerce customer churn prediction by gradient boosted trees. En *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)* (pp. 055–059). IEEE.

Sabbeh, S. F. (2018). Machine-learning techniques for customer retention: A comparative study. *International Journal of Advanced Computer Science and Applications, 9*(2), 273.

Schouten, A. P., Janssen, L., & Verspaget, M. (2019). Celebrity vs. influencer endorsements in advertising: The role of identification, credibility, and product-endorser fit. *International Journal of Advertising*. https://doi.org/10.1080/02650487.2019.1634898

Scikit-learn developers. (n.d.). *sklearn.model_selection.GridSearchCV*. Scikit-learn. Recuperado el 24 de abril de 2025, de https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing, 5*(4), 13–22.

Shobana, J., Gangadhar, C., Arora, R. K., Renjith, P. N., Bamini, J., & Chincholkar, Y. D. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors, 27*, 100728. https://doi.org/10.1016/j.measen.2023.100728

Statista. (2024). *Influencer marketing market size worldwide from 2015 to 2025* [Gráfico]. Statista. https://www.statista.com/statistics/1092819/global-influencer-market-size/

Vafeiadis, T., Diamantaras, K., Chatzisavvas, K. C., & otros autores. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory, 55*, 273–285. https://doi.org/10.1016/j.simpat.2015.03.003

White, S. A., & Miers, D. (2008). *BPMN modeling and reference guide: Understanding and using BPMN*. Future Strategies Inc.

Xia, G., & He, Q. (2018). The research of online shopping customer churn prediction based on integrated learning. En *2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018)* (pp. 259–267). Atlantis Press.

Yanfang, Q., & Chen, L. (2017). Research on e-commerce user churn prediction based on logistic regression. En *2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (pp. 87–91). IEEE.

Anexo A. Consulta SQL. Extracción del conjunto de datos.

```
links_with_users AS (
    SELECT
        pl.link_id,
        pl.creator_id,
        pl.source_id,
        pl.link_date,
        u.id AS user_id,
        u.locked_at IS NOT NULL AS is_locked,
        u.created_at AS register_date,
        fu.id IS NOT NULL AS is_fix
    FROM product_links_de AS pl
    JOIN rawdata.pg_prod_public.users AS u
        ON pl.creator_id = u.id
    LEFT JOIN rawdata.pg_prod_public.fix_users AS fu
        ON fu.user_id = u.id
    WHERE u.role_id = 0
      AND fu.id IS NULL
      AND u.locked_at IS NULL
)
SELECT
    lwu.user_id,
    lwu.register_date,
    lwu.link_id,
    lwu.link_date,
    lwu.source_id,
    COALESCE(SUM(t.GMV), 0) AS GMV,
    COALESCE(SUM(t.Provision), 0) AS Provision,
    COALESCE(SUM(t.No_Transactions), 0) AS No_Transactions,
    COALESCE(SUM(c.Commission), 0) AS Commission,
FROM links_with_users AS lwu
JOIN commission_agg AS c
    ON lwu.link_id = c.product_link_id
LEFT JOIN transactions_agg AS t
    ON lwu.link_id = t.product_link_id
GROUP BY
    lwu.user_id,
    lwu.register_date,
    lwu.link_id,
    lwu.link_date,
    lwu.source_id
```

```
WITH product_links_de AS (
    SELECT
        pl.id AS link_id,
        pl.creator_id,
        pl.created_at AS link_date,
        pl.source_id
    FROM rawdata_copy.pg_prod_public.product_links_clone AS pl
    WHERE pl.created_at BETWEEN '2023-01-01' AND '2025-04-01'
      AND pl.country = 'de'
),
transactions_agg AS (
    SELECT
        t.product_link_id,
        SUM(t.gross_basket_eur) / 100.0 AS GMV,
        SUM(t.gross_provision_eur) / 100.0 AS Provision,
        COUNT(DISTINCT t.transaction_id) AS No_Transactions
    FROM mart.stylink.transactions_processed AS t
    GROUP BY t.product_link_id
),
commission_agg AS (
    SELECT
        s.product_link_id,
        SUM(s.confirmed_commission_eur) / 100.0 AS Commission
    FROM stage.pg_prod_public.stats_current AS s
    GROUP BY s.product_link_id
),
```

Anexo B. Árbol de decisión ampliado.