UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

# DESARROLLO DE UN MODELO DE RIESGO CREDITICIO PARA CLEO CHILE

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

**GONZALO ALONSO DONOSO SALAS**

PROFESORA GUÍA:
Loreto Martínez Giménez

MIEMBROS DE LA COMISIÓN:
Sabino Aguad Merlez
Alejandra Puente Chandía

SANTIAGO DE CHILE
2024

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: GONZALO ALONSO DONOSO SALAS
FECHA: 2024
PROF. GUÍA: LORETO MARTÍNEZ GIMÉNEZ

# DESARROLLO DE UN MODELO DE RIESGO CREDITICIO PARA CLEO CHILE

El presente informe documenta el proyecto de título que aborda el desarrollo de un nuevo modelo de riesgo crediticio para Cleo Chile, una empresa *Fintech* especializada en servicios de compra en cuotas en línea (BNPL, por sus siglas en inglés).

Durante la última década, la industria *Fintech* en Chile ha crecido un 29 %, intensificando la competencia en el mercado con la entrada de grandes compañías. Cleo, como parte de este sector, ha operado hasta ahora con un modelo de riesgo que evalúa la capacidad de pago de los clientes. Sin embargo, dicho modelo ha perdido precisión desde su implementación, resultando en un incremento en la tasa de no pago y generando pérdidas que han afectado negativamente el desempeño financiero de la compañía.

Ante esta problemática, el objetivo del proyecto es diseñar un nuevo modelo de riesgo que ofrezca mayor precisión en la predicción del comportamiento de pago de los usuarios, reduciendo así la tasa de error y mejorando la utilidad generada por el servicio BNPL.

El desarrollo del modelo se realizó siguiendo la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), que abarca etapas iterativas de comprensión del negocio, preparación de datos, modelado, evaluación y despliegue. Esta metodología permitió ajustar y optimizar continuamente el modelo a través de iteraciones, utilizando datos históricos proporcionados por Cleo. Se emplearon modelos de clasificación supervisada reconocidos en la literatura por su alta precisión en la predicción de riesgos crediticios, específicamente los algoritmos Random Forest, XGBoost y LightGBM.

Durante el modelado, se identificaron variables clave que influyen en el comportamiento de pago, como edad, monto de compra, deuda, ingresos mensuales y saldo en cuenta. Estas variables se integraron en los modelos para mejorar su rendimiento. Los modelos desarrollados fueron evaluados utilizando métricas específicas, como la utilidad esperada, la tasa de dinero prestado a falsos positivos y el valor AUC (Area Under the Curve), lo que permitió seleccionar el modelo más eficiente. El nuevo modelo demostró una mejora significativa, reduciendo la tasa de no pago en un 90 % y aumentando la precisión en la predicción de pagadores en un 50 %, cumpliendo así con el objetivo del proyecto.

La implementación del nuevo modelo podría contribuir a la sostenibilidad y el crecimiento futuro de Cleo Chile. Al reducir el riesgo asociado a las compras a crédito, la empresa podría ofrecer un mayor cupo de compra a los usuarios, mejorando su experiencia y aumentando su fidelización. Además, Cleo podría mejorar su posición competitiva en el mercado de las *Fintech*, con capacidad para adaptarse en un entorno en constante cambio. Por último, la metodología y los resultados obtenidos en este proyecto sirven como base para futuros desarrollos en la optimización de modelos de riesgo en la industria financiera.

*La vida es una lenteja,*
*o la tomas, o la dejas.*

***Saludos***

# Agradecimientos

La realización de este documento marca un antes y un después, ya que pone fin a una época de altos y bajos, llamada vida universitaria, para dar paso a un ciclo nuevo, la vida laboral. Algunos recordarán esta época por las conexiones realizadas, por los amigos que hicieron en el camino o incluso por logros académicos conseguidos en este transcurso temporal, mientras que otros quizás recordarán con un mayor pesar esta etapa de sus vidas y con algunos tonos no tan brillantes.

Por mi parte, siento que el ámbito académico de exigencia pone en juego las capacidades de la persona sobre cómo llevar su estancia en la universidad, lo que va moldeando cómo vive esta experiencia. En mis primeros recuerdos del inicio de la universidad solo surgen imágenes borrosas y un sentimiento de vacío, el cual se retroalimentaba al pensar que lo único necesario para una buena estancia radicaba en tener un buen rendimiento, dejando de lado aspectos como mi círculo social o a veces mi propio cuidado.

Primaveras han pasado desde ese vago recuerdo, desde esas épocas en las que trasnochaba y me exigía día tras día solo por alcanzar un número que demostrara mi valía dentro de este ambiente. Agradezco no haber seguido por este camino, ya que era solitario e insostenible en el tiempo. Agradezco a Su por mostrarme su apoyo en un momento en el que ni siquiera hubiera pensado en necesitarlo, y también a Seba y a Laura de las abejitas, por haberme acompañado cuando recién pude comenzar a abrirme con los demás en la U, y seguir haciéndolo aún, lo cual recordaré con una especial estima. Y agradezco a mis amigos de liceo, que a pesar de no estar tan cerca como desearía, sé que siempre serán un gran apoyo para mí.

Así, mis últimos años de universidad serán recordados como una mejoría para mí mismo, una fase en la que comencé a salir con los demás, a almorzar acompañado, a distenderme, e incluso a vivir más allá del contexto académico, cosa impensada por mí en los primeros años. Este último tiempo también será el que en vez de enfocarme en conseguir una calificación excelente, me enfoqué en ser una excelente persona, para mí y para los demás.

Claro que este cambio y todo esto no sería posible sin mi familia, que incondicionalmente mostró su apoyo y cuidado hacia mi persona. Doy gracias a mis abuelos, que siempre demostraron su cuidado y su querer; a Mario, por estar incondicionalmente, a pesar de mi poca expresividad; a mi hermano, por siempre guiarme y estar abierto a escucharme independientemente de lo lejos que estemos; a mi papá, que a pesar de que ya no te pueda ver, aún siento que me acompañas; y a mi mamá, que simplemente debería darle las gracias por todo, me siento orgulloso de ser su hijo.

También es necesario agradecer a mi equipo de Cleo que siempre se mostró abierto y con la mejor de las disposiciones para lograr completar este proyecto, además de disponer de su tiempo para que aprendiera, y gracias a mis profesores guía por su retroalimentación para conseguir un entregable óptimo y por mostrar su disposición a conseguir esto.

Y no puedo olvidar a mis amigos de 4 patas, Apolo y Hera, que me demostraban día a día uno de los sentimientos más puros que algún ser puede llegar a imaginar, llegando a ser más que solo mascotas. Además, hicieron el papel de una fuerte barrera que en momentos evitaba mi descenso a la locura, mi lugar seguro, el que me desconectaba de un mundo en ocasiones oscuro.

Finalmente, y es algo que pocas veces se le da importancia, gracias Gonzalo, a lo largo de los años pudiste dejar de estar solamente en tu zona de confort, la que en cierta medida no era totalmente saludable para ti, aprendiste a pedir ayuda, a pesar de ir en contra de tu naturaleza, y aprendiste que la vida no solo se trata de qué tan bien te ven las otras personas, sino de qué tan bien te ves a ti mismo.

Quisiste cambiar, quisiste enfrentar el ahora, por muy intenso y doloroso que a veces sea, quisiste sentir, y sobre todo, quisiste dejar de pensar en que el futuro sería algo mejor, ya que eso no te permitía ver lo que estaba pasando en el presente. Tal como dijo John Lennon: "La vida es eso que sucede mientras estás ocupado haciendo otros planes"[1]. A veces uno evita esto, se olvida o lo ignora, se disocia, pero puedo asegurar que nunca me había sentido más vivo, y que cada parte de mi vida me trajo hasta este momento.

[1] Extracto de Beautiful Boy (Darling Boy).

# Tabla de Contenido

# Índice de Tablas

# Índice de Ilustraciones

# 1. Antecedentes Generales

## 1.1. Descripción de la organización

La presente memoria se realiza en el marco de un servicio de compra en cuotas en línea en Cleo Chile SpA, una organización que forma parte del segmento de empresas reconocidas como *Fintech*[1]. Según la Comisión Nacional del Mercado de Valores[2], las *Fintech* se definen como entidades que utilizan la innovación y desarrollos tecnológicos para ofrecer servicios financieros.

Para comprender la situación de la compañía, se comenzará por analizar el entorno que la rodea, iniciando con un análisis PESTEL. Posteriormente, se visualizará el mercado actual y la competencia presente, seguido de las proyecciones de esta industria. Finalmente, se describirá con más detalle a la organización y se realizará un análisis FODA para comprender las variables internas de la empresa.

## 1.2. Análisis del entorno

### 1.2.1. Análisis PESTEL

Para comenzar el análisis del entorno en el que opera Cleo, se llevará a cabo un análisis PESTEL de la industria de las *Fintech* en Chile, examinando los aspectos políticos, económicos, sociales, tecnológicos, ecológicos y legales.

- **Político:** Dada la relevancia del entorno político para cualquier industria, especialmente en términos de estabilidad política y económica, resulta relevante considerar que los indicadores proporcionados por Cadem indiquen que la percepción de los ciudadanos sobre el país se haya estancado o empeorado en los últimos dos años[3].
- **Económico:** En Chile, el panorama económico parece ser auspicioso para las empresas, ya que según datos oficiales del gobierno[4], el Índice Mensual de Actividad Económica ha registrado un incremento del 3,5 % en el mes de abril de 2024, sobre las expectativas anuales de un 2,7 %.
- **Social:** Se observa un cambio significativo en el comportamiento de compra de los consumidores luego de la pandemia de 2020, donde se prefiere la compra presencial en lugar de la remota, o bien, se opta por tiendas con ambos métodos de venta. De hecho, según cifras de Instore Media, el 82 % de las decisiones de compra se toman de forma presencial, y el 60 % de los nuevos productos se conocen en tienda[5] (Ponasso, 2024).

[1] Por sus siglas en inglés referentes a *Finance and Technology*.
[2] Para más información: https://www.cnmv.es/DocPortal/Publicaciones/Fichas/GR03_Fintech.pdf
[3] Para más información: https://cadem.cl/estudios/84-piensa-que-el-pais-esta-peor-que-hace-dos-anos-en-crimen-organizado-y-82-en-delincuencia/
[4] Para más información: https://www.hacienda.cl/noticias-y-eventos/noticias/ministro-marcel-sobre-imacec-de-abril-de-3-5-nos-confirma-que-la-economia-se
[5] Para más información: https://portaleduca.cl/relevancia-de-la-tienda-fisica-en-el-retail-para-esta-vuelta-a-clases-segun-cencosud-media/

- **Tecnológico:** A pesar de que el comercio electrónico registró una caída durante tres años consecutivos, según información de la Cámara de Comercio, para el año 2024 el crecimiento se perfila en un 8 %[6]. Por otro lado, según Mena Álamos (2023), *Ecosystem Product Owner* de BCI Labs, las billeteras digitales y los pagos móviles, a través de dispositivos inteligentes, tomarán el primer lugar en las preferencias de los consumidores. Además, se prevé que en los próximos 10 años la tokenización[7], el uso de la IA y la mejora de la experiencia del cliente serán las tendencias dominantes.

- **Ecológico:** Según menciona Mercado (2022) de El País, las *Fintech* representan un ejemplo ideal para otras industrias en términos ambientales. Su enfoque en la innovación y la creatividad les permite ser sostenibles y capaces de adoptar tecnologías cada vez más eficientes con un menor impacto ambiental, como se destaca también en la revista *International Journal of Computers & Technology* (Tavor et al., 2013).

- **Legal:** En Chile, la regulación del mercado de las *Fintech* recae principalmente en dos instituciones: el Banco Central de Chile, encargado de fomentar la estabilidad y eficacia del sistema financiero, y la Comisión para el Mercado Financiero (CMF), responsable de promover el desarrollo y la estabilidad del mercado financiero[8].

  La regulación de las *Fintech* es relativamente nueva, debido a que su ley mandatoria fue publicada el 4 de enero del año 2023 ("Promueve la competencia e inclusión financiera a través de la innovación y tecnología en la prestación de servicios financieros, Ley Fintec. Ley Nº21.521", 2024). Además de la ley, la CMF está facultada para emitir normas para regular a las empresas que se adscriben a esta industria[9].

Considerando el análisis previo, se puede afirmar que la industria mantiene una posición favorable en lo que se refiere a variables macroeconómicas como el crecimiento del país y la inflación, así como las tendencias ecológicas y del mercado. Además, surge una oportunidad de estas nuevas preferencias, referente a absorber parte de la demanda creciente del público que realiza sus compras de manera presencial. Es por esto que, según una investigación de Carmila, un 70 % de las marcas digitales prevé abrir un local en el futuro (Ponasso, 2024).

Por otro lado, los factores tecnológicos indican que Cleo deberá incorporar tecnologías como la tokenización, con el objetivo de reforzar la seguridad, y los pagos móviles en el futuro cercano, dado su alineamiento con los servicios actuales.

Finalmente, los aspectos políticos y legales requieren cautela, ya que son variables inciertas en este momento, sujetas al desempeño del gobierno y a la adopción de nuevas regulaciones por parte del parlamento. La ley *Fintech* en el segundo cuarto del 2024 sigue convocando grupos consultivos, lo que advierte un cambio constante y que genera incertidumbre.

### 1.2.2. Situación actual

Según Berstein Jáuregui (2024), entre los años 2019 y 2023, la industria de las *Fintech* chilenas ha experimentado un crecimiento anual de aproximadamente un 29,5 %,

[6] Para más información: https://www.ccs.cl/2024/04/25/camara-de-comercio-de-santiago-preve-crecimiento-del-8-en-el-comercio-electronico-durante-2024/

[7] Encriptación dinámica de los datos del cliente durante una transacción (Alameda, 2019).

[8] Para más información:https://www.cmfchile.cl/portal/principal/613/w3-propertyvalue-25539.html

[9] Para más información: https://www.cmfchile.cl/portal/prensa/615/articles-77135_doc_pdf.pdf

alcanzando un total de 300 empresas para la fecha del informe, como se detalla en el Anexo A.1. Además, otras 78 empresas extranjeras han ingresado al mercado nacional, según lo reportado por el Fondo Monetario Internacional (FMI)[10].

Este crecimiento se refleja en el aumento del acceso a cuentas bancarias en Chile, así como en la búsqueda de medios de pagos más eficientes y rápidos. En el año 2021, un 85 % de los chilenos poseían una cuenta bancaria y habían utilizado métodos de pago digitales, en comparación con un 50 % entre los años 2011 y 2014, según informa el FMI[10]. Además, según un informe del comisionado de la CMF, Kevin Cowan (Cowan, 2022), los pagos con tarjeta y las transferencias han experimentado un crecimiento, mientras que el uso de métodos como cheques y cajeros automáticos ha disminuido (Ver Anexo A.2)

Para realizar un análisis más detallado de la industria, es necesario dividirla en segmentos específicos, ya que la definición general de las *Fintech* abarca un espectro amplio. Es por esto que la CMF (Berstein Jáuregui, 2024) ha identificado 11 segmentos distintos basados en los servicios que ofrecen las empresas de este sector, como se muestra en la Figura 1.1.

Dentro de esta clasificación, Cleo se encuentra principalmente en el segmento de Pagos y Remesas, aunque debido a la naturaleza de su servicio, también podría considerarse en Préstamos. Por lo tanto, solo se analizarán estas dos categorías. En estos segmentos se encuentran plataformas reconocidas en el ámbito digital, como CopecPay, Klap y Tapp[11].

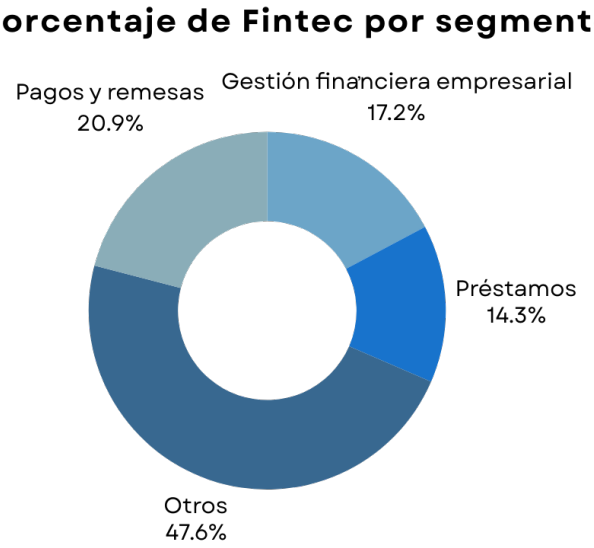


Figura 1.1: Porcentaje de Fintec por segmento. Fuente: Adaptado de *Ley Fintec Desafíos de la implementación.*- CMF.

### 1.2.3. Competencia

Como se mencionó en la Sección 1.1, el servicio a analizar corresponde a un método de pago que permite a los usuarios realizar compras en cuotas, llamado *Buy Now Pay Later*. Se podría considerar como competidores a cualquier emisor de tarjetas de débito o crédito,

[10] Para más información: https://www.elibrary.imf.org/view/journals/002/2024/042/article-A003-en.xml
[11] Estas firmas forman parte de billeteras digitales y operadores de tarjetas de pago.

e incluso podría incluirse el pago en efectivo. Sin embargo, dado que el servicio ofrecido apunta a personas que no poseen tarjetas de crédito, o que tengan algún problema para utilizarlas, y requieren comprar en cuotas, los medios anteriormente mencionados podrían ser parte de una competencia indirecta.

En este sentido, empresas que ofrecen pagos en cuotas, como la argentina Wibond[12] o la chilena Ventipay[13], que cuentan con más de 40 y 1000 comercios respectivamente, según se visualiza en sus páginas, podrían considerarse competidores directos de Cleo. Sin embargo, según Tejeda (2024), Cleo se destaca como una empresa líder en este servicio para Chile al analizar el Informe de pagos globales 2023 de Wordpay. Además, Cleo también presenta comisiones en promedio más bajas que Wibond y Ventipay para los comercios, como se observa en la Tabla A.1. Empresas reconocidas como MACH del banco BCI han comenzado a implementar el pago en cuotas para algunas cuentas seleccionadas[14]. No obstante, aún no se han instalado en Chile plataformas reconocidas a nivel mundial como Klarna, PayPal Credit, Affirm y AfterPay, las cuales podrían tener un impacto significativo en este mercado en el futuro[15].

Considerando lo anterior, es evidente que el mercado de BNPL está evolucionando desde uno de nicho hacia uno más amplio, especialmente si firmas como MACH están tratando de ingresar. Esto indica que la competitividad del sector se encuentra en aumento, por lo que Cleo necesita mantenerse innovadora y eficiente para seguir siendo un actor relevante.

### 1.2.4. Proyecciones de la industria

Según proyecciones recientes de Cardozo (2023) y de De Losada (2024), se espera que la industria *Fintech* experimente un crecimiento significativo en el país en los próximos años. Esto se ve respaldado por las conclusiones obtenidas en el informe realizado por Finnovista en colaboración con Visa en el año 2023 sobre la evolución y el estado de la industria[16].

En el caso específico de los servicios ofrecidos por Cleo, el panorama futuro también luce prometedor. Para el mercado de servicios de BNPL, según datos de Moreno (2024) y de Pescio (2022) de Diario Financiero, se anticipa un crecimiento anual compuesto de entre un 26 % a un 29 % para fines de la década, entre el 2028 y el 2030. Estos datos sobre el crecimiento de la industria y del servicio pueden llegar a ser auspiciosos para las empresas del sector, como Cleo, ya que se espera que la compañía continúe creciendo en los años venideros, lo que correspondería a un escenario favorable. No obstante, también sería esperable que lo hicieran las compañías competidoras, por lo que es esencial tener cautela en este aspecto.

## 1.3. Análisis interno

La compañía llega a Chile desde Estocolmo en el año 2013 bajo el nombre de SweetPay, y posteriormente cambia su denominación a Cleo en 2018. Este nacimiento empresarial ocurrió apenas un año después de que Khipu y Cumplo, las primeras *Fintech* en Chile, se establecieran en el mercado en el año 2017 (Christiansen, 2022). Tomando esta referencia,

[12] Para más información: https://wibond.cl/shops
[13] Para más información: https://ventipay.com/personas/
[14] Para más información: https://ayuda.somosmach.com/hc/es/articles/12164831090317
[15] Para más información: https://www.fabrick.com/es-es/recursos/blog/que-es-buy-now-pay-later/
[16] Para más información: https://www.finnovista.com/radar/chile2023/

se podría argumentar que Cleo posee una ventaja competitiva en el conocimiento del consumidor chileno en comparación con nuevas empresas del sector.

### 1.3.1. Productos

En 2024, Cleo ofrece 2 productos principales al mercado, los cuales son:

- **Transfer:** Servicio integrado de pago y retiro de dinero. Tiene como principales clientes a los servicios de *Igaming*[17], que corresponden principalmente a casinos online.

  Además de ofrecer los servicios de pago y retiro, Cleo puede proporcionar información de los usuarios a los comercios si es requerida, con el objetivo de que estos últimos puedan conocer más a sus usuarios y ofrecer un servicio más personalizado.

- **Buy Now Pay Later (BNPL):** Método de pago en línea que permite al usuario pagar en cuotas con bajos intereses, sin necesidad de poseer una tarjeta de crédito. Este producto forma parte de una solución para la inclusión financiera de los consumidores, permitiéndoles tener una mayor capacidad de compra y de pago a plazo. El modelo de negocio de Cleo innova al integrar el riesgo individual, al ofrecer opciones de pago a plazos. Esto se logra mediante la generación de ingresos a través de comisiones cobradas al comercio y tasas de interés aplicadas al comprador en cada cuota de pago. Este modelo se basa principalmente en las siguientes fases:

  – Compra: El usuario realiza su orden a través de su comercio de preferencia y selecciona a Cleo como método de pago.
  – Evaluación de riesgo: Con base en información obtenida con el consentimiento de la persona, referente a datos bancarios y de la compra, se realiza una evaluación de la capacidad crediticia de la persona con el objetivo de obtener una probabilidad de pago para saber si aceptar o denegar la compra. Actualmente, es utilizado un modelo diseñado el año 2023 con el nombre de V4.
  – Postventa: Si la compra fue aprobada, el usuario recibe su compra y se genera el cargo referente a las cuotas seleccionadas en el plazo establecido.

Es necesario notar que en términos de ingresos para la compañía, Transfer abarca más del 90 % y BNPL correspondería al resto del porcentaje para completar el 100 % según datos internos.

### 1.3.2. Misión y visión

Según reuniones con el equipo de Cleo, la misión corresponde a *Transformar la industria Fintech al proporcionar soluciones de pagos centradas en el usuario. Nos comprometemos a crear soluciones disruptivas que impulsen el cambio y brinden un impacto duradero en la vida de las personas.*

La visión señala: *Convertirnos en líderes regionales en soluciones de pagos centradas en el usuario, ofreciendo una experiencia excepcional y superando las expectativas de nuestros clientes. Nos esforzamos por ser pioneros en innovación, establecer nuevos estándares de excelencia y ser reconocidos como agentes de cambio en la industria Fintech, transformando la forma en que las personas realizan transacciones financieras.*

[17] Apuestas en internet sobre el resultado de un juego de azar (Marmuzevich, 2023). Abarcando casinos, apuestas deportivas, juegos de cartas, entre otros.

### 1.3.3. FODA

Luego de indagar en el mercado en el que se desenvuelve Cleo, es necesario profundizar en los aspectos de la empresa, para lo cual se realiza un análisis FODA con el objetivo de diagnosticar la situación estratégica de la organización.


Figura 1.2: Análisis FODA. Fuente: *Elaboración propia*

El análisis FODA, representado en la Figura 1.2 y detallado en el Anexo A.3, muestra que Cleo, a pesar de ser una empresa emergente en el dinámico sector Fintech, posee robustas fortalezas y prometedoras oportunidades que puede explotar para consolidar y expandir su presencia en el mercado. Es imperativo que Cleo capitalice su equipo humano y su capacidad para innovar y adaptarse rápidamente a las nuevas tendencias tecnológicas para mantenerse competitivo. Asimismo, no debe subestimar las debilidades y amenazas identificadas, como su tamaño reducido, la volatilidad regulatoria de Chile, y en especial, la alta tasa de no pago que muestra su servicio.

Algunas recomendaciones para aumentar su cuota de mercado, basadas en el análisis anterior, corresponderían a que Cleo debería enfocarse en tres estrategias principales: retener y continuar desarrollando su talento humano, por ejemplo a través de cursos y talleres; explorar y adoptar nuevas tecnologías y modelos de negocio que diversifiquen sus fuentes de ingresos y permitan disminuir el no pago, lo que impulsaría la sostenibilidad a largo plazo; y fortalecer su capacidad para adaptarse a los marcos regulatorios. Implementando estas estrategias, Cleo no solo podrá mejorar su competitividad, sino también asegurar una base sólida para su crecimiento.

## 1.4. Rol del estudiante

El estudiante responsable de esta memoria desempeña el rol de analista de datos en el proyecto de diseño de un nuevo modelo de riesgo. Este rol se encuentra bajo la supervisión del *product owner* del proyecto y trabaja en estrecha colaboración con el equipo de nuevos productos, asegurando una alineación efectiva de los objetivos del proyecto con las necesidades y prioridades del negocio. Las actividades del estudiante incluyen colaborar estrechamente con el equipo encargado de BNPL para discutir las funcionalidades deseadas en el nuevo modelo y los cambios necesarios para mejorar la elegibilidad de los usuarios. Además, el estudiante se dedica a investigar posibles mejoras basadas en la literatura disponible, implementar las iteraciones necesarias utilizando Python, el lenguaje definido por la organización, y evaluar cualquier sesgo presente en el modelo.

# 2. Justificación del tema

## 2.1. Descripción del problema

Considerando el desempeño de los últimos meses en el servicio BNPL de Cleo, se ha observado un aumento constante en los niveles de *default*, o no pago, por parte los usuarios, como se muestra en la Figura 2.1. En esta representación gráfica, el no pago general refleja a aquellos usuarios que efectivamente no realizan los pagos, mientras que el no pago del modelo actual, mencionado en la Sección 1.3.1, corresponde a las predicciones realizadas por este. Sería razonable esperar que si los niveles de incumplimiento previstos por el modelo se mantuvieran bajos, los del no pago general también lo harían. Por lo tanto, el problema radica en que el no pago real está en aumento porque el incumplimiento predicho también lo está, lo que impacta negativamente en la rentabilidad del servicio, como se aprecia en el Anexo B.1, dónde existen meses en los que la tasa de no pago supera el umbral de rentabilidad, generando pérdidas.

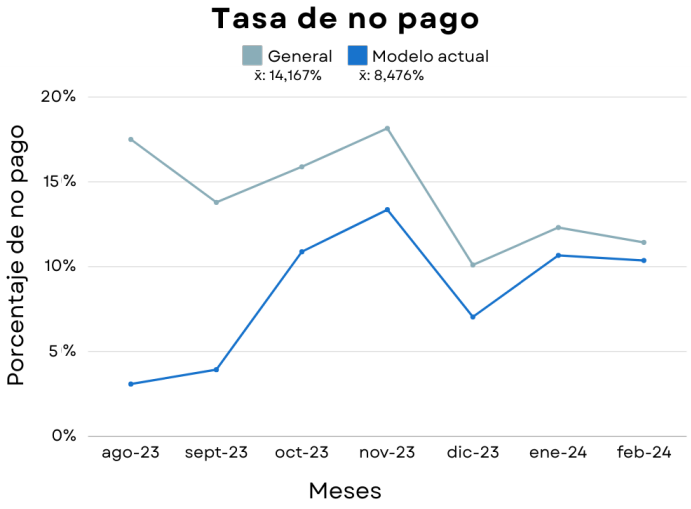


Figura 2.1: Gráfica de tasas de no pago en el último período. Fuente: *Elaboración propia*

Dado que BNPL es un método de pago basado en la extensión de crédito para compras específicas y depende de la confianza en el pago por parte del usuario, es crítico que los niveles de incumplimiento aumenten, ya que esto afecta directamente la rentabilidad de la organización. Si esta tendencia persiste, podría convertirse en un problema significativo y difícil de resolver, afectando negativamente tanto al servicio de compras en línea como a sus planes futuros.

Para analizar el problema, se ha elaborado un diagrama de Ishikawa que se presenta en la Figura 2.2. Esta herramienta es útil para identificar las posibles causas subyacentes del problema, centrándose en los factores generales que podrían estar contribuyendo a él y que,

a su vez, podrían ser influenciados por otros factores más específicos.

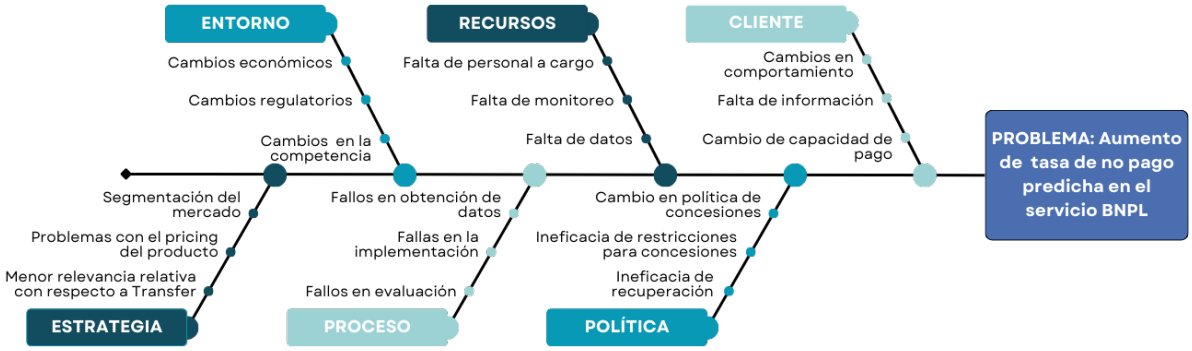


Figura 2.2: Diagrama de Ishikawa. Fuente: *Elaboración propia*

Durante el análisis, se han identificado seis factores que podrían estar influyendo en el problema mencionado, los cuales se detallan a continuación:

- **Estrategia:** Las variables estratégicas que pueden estar afectando el servicio BNPL incluyen la segmentación del mercado, el precio del servicio en comparación con la competencia y la menor importancia que posiblemente se le da a BNPL dentro de la compañía en términos de ingresos, dado que Transfer genera mayores ingresos, como se mencionó en la Sección 1.3.1.

- **Entorno:** Factores externos como el aumento del desempleo, que alcanzó un 8,4 % entre noviembre de 2023 y enero de 2024[18], podrían estar contribuyendo al fenómeno del impago por parte de los usuarios. Además, cambios legislativos, como la ley de protección de datos personales (Olmos, 2024), así como la evolución en la competencia, también podrían estar ejerciendo influencia.

- **Proceso:** La operación del servicio BNPL podría estar comprometida por errores en la obtención de datos de usuarios y la dificultad para acceder a información sensible. Además, según el equipo, el modelo actual fue desarrollado apresuradamente para mejorar problemas anteriores, por lo que no sería totalmente óptimo, lo que puede afectar la precisión de las predicciones.

- **Recursos:** La falta de un equipo dedicado al análisis de riesgos y la incapacidad para monitorear adecuadamente las aprobaciones y registros de compra pueden llevar a una detección tardía de fallos relevantes. Además, disponer de más datos de los usuarios podría mejorar las predicciones del modelo.

- **Políticas:** Los cambios en las políticas para determinar la elegibilidad de los créditos y las políticas de recuperación pueden estar contribuyendo al impago de los usuarios.

- **Cliente:** Los usuarios han experimentado cambios en su comportamiento, como se mencionó en el análisis PESTEL, y en su capacidad de pago, ya que como menciona el economista Víctor Salas, debido a la inflación sobre el 3 %, los salarios reales tenderán a disminuir (Martinez, 2024). Además, algunos pueden no estar dispuestos a autorizar

[18] Para más información: https://elpais.com/chile/2024-02-28/el-desempleo-en-chile-llega-a-84-presionado-por-al-alza-de-la-fuerza-de-trabajo.html

el uso de su información, lo que afecta los perfiles generados y la efectividad del servicio.

## 2.2. Descripción y justificación del proyecto

Considerando la descripción y análisis detallados en la sección anterior sobre el problema que enfrenta actualmente Cleo, se puede observar que los factores influyentes provienen tanto de variables externas a la organización, como variables macroeconómicas, variables internas, y el funcionamiento del modelo de riesgo dentro del servicio. Si bien aspectos del entorno, como los niveles de desempleo, regulaciones gubernamentales y el comportamiento del cliente, son áreas que no pueden ser controladas por la empresa, es posible abordarlas mediante estrategias de gestión.

Para enfrentar el problema, se opta por una solución más eficiente que pueda abarcar una mayor cantidad de factores que podrían estar provocando el problema. Esto implica intervenir en el diseño y mejora del modelo actual utilizado para la aprobación de compras en BNPL. La elección de esta solución se basa en la capacidad de la empresa para gestionar internamente el proceso de diseño y optimización del modelo de riesgo, a diferencia de factores externos sobre los cuales no tiene control directo.

Además, el rendimiento del modelo se puede medir en un período corto de tiempo con una métrica definida, como la tasa de error de predicción, a diferencia de otras estrategias que podrían incluir un proceso de cambio organizacional, el cual podría tomar entre 1 a 5 años para mostrar resultados[19]. No obstante, estos factores externos no pueden ser ignorados, por lo que en el proceso de rediseño del modelo de riesgo también se deben considerar estrategias para mitigar sus efectos, como la implementación de medidas flexibles que se adapten a cambios en el entorno económico o regulatorio.

Cabe destacar que, aproximadamente, cada venta realizada por Cleo genera una ganancia correspondiente al 10 % del monto, considerando las comisiones cobradas al usuario y al comercio. Tomando en cuenta esto y suponiendo ventas de igual valor, si una persona no realiza el pago, se necesitarían en promedio otras 10 ventas con usuarios que efectivamente paguen para compensar esa pérdida. Esto refleja la importancia de predecir correctamente a qué personas se les puede conceder un préstamo y a cuáles no. Esta relevancia se ve aumentada por la competitividad del mercado en el que Cleo opera.

El proyecto busca abordar los principales factores identificados en la Figura 2.2. El rediseño y optimización del modelo actual resolverá posibles fallos en la operación actual, permitiendo una mejor captura de la información del usuario para comprender su comportamiento. Además, el nuevo modelo incluirá una redefinición de las políticas de elegibilidad de préstamos, lo que modificará en gran medida el modo de otorgar los créditos de compra. Estas políticas revisadas serán diseñadas considerando la capacidad de adaptación a diferentes escenarios externos y la mitigación de riesgos asociados.

[19] Para más información: https://albertogarciacarro.com/faq/cuanto-suele-durar-un-proceso-de-gestion-del-cambio/

# 3. Objetivos

## 3.1. Objetivo general

Desarrollar un modelo de riesgo crediticio, para la empresa Cleo Chile, que presente una tasa de error menor que la del modelo actual.

En esta memoria el error del modelo será evaluado a través de la métrica de la utilidad del modelo, la tasa de dinero prestada a los falsos positivos[20] (fp) y del valor *Area Under the Curve* (AUC)[21]. El plazo para el desarrollo del proyecto corresponde a 4 meses.

## 3.2. Objetivos específicos

- Analizar los datos de los usuarios del servicio BNPL para comprender su comportamiento y buscar posibles tendencias entre buenos y malos pagadores.
- Realizar una revisión detallada de la literatura académica sobre modelos de predicción para fundamentar las decisiones de diseño del nuevo modelo.
- Diseñar y desarrollar múltiples modelos de riesgo utilizando diversas técnicas de aprendizaje de máquinas para determinar cuál ofrece una mejor precisión y eficacia predictiva.
- Preparar y validar el modelo propuesto con el equipo de trabajo, asegurando su integración fluida con los sistemas existentes.

[20] Un falso positivo se refiere a predecir un valor como positivo, cuando en realidad sería negativo (Mankad, 2020).
[21] Métrica que indica la capacidad del modelo para distinguir entre clases (Narkhede, 2018).

# 4. Alcances

En relación con los objetivos anteriores, se definen los límites del proyecto, considerando las restricciones técnicas, de recursos y temporales. Los alcances del trabajo son los siguientes:

- El modelo se desarrollará exclusivamente para el servicio BNPL de Cleo. Aunque podría extenderse a otros servicios en desarrollo, esta opción no se considerará en el proyecto actual debido a las limitaciones de tiempo. La extensión requeriría una fase adicional de análisis y desarrollo que excede el límite temporal dado para el trabajo.

- Los datos utilizados para el modelado estarán limitados a aquellos que la organización puede obtener hasta la fecha, tanto interna como externamente. Esto incluye datos históricos disponibles y aquellos accesibles a través de fuentes externas. La limitación en los datos se debe a la disponibilidad y al tiempo necesario para su recopilación. Se asume que un conjunto de datos más amplio y detallado podría mejorar significativamente la precisión del modelo.

- Debido a las restricciones en la capacidad de procesamiento de la computadora utilizada para el desarrollo del proyecto, se optará por modelos de complejidad moderada. Así, métodos avanzados como redes neuronales, que requieren altos recursos computacionales (Donges, 2023), quedan fuera del trabajo.

- Finalmente, el proyecto abarcará el traspaso de la información del modelo, su documentación detallada y su paquete de datos al equipo de BNPL. La implementación directa del modelo en la página web de Cleo no está incluida en el alcance del trabajo debido a las limitaciones temporales y a la falta de experiencia técnica del memorista.

El enfoque en el nuevo diseño de modelo se justifica porque soluciones alternativas, como una reestructuración completa del proceso de compra, implicarían un mayor tiempo de desarrollo y desafíos significativos para medir sus resultados de manera efectiva. Por ejemplo, una reestructuración del proceso podría afectar diversos aspectos operativos y no proporcionar resultados inmediatos ni métricas claras. Dado que el nuevo modelo puede desarrollarse dentro del plazo establecido y medirse mediante métricas específicas como la utilidad entregada, la tasa de dinero prestado a los falsos positivos y el valor AUC, se considera la opción más adecuada. Además, el modelo se desarrollará de forma interna y controlada, permitiendo ajustes y evaluaciones sin impactar la operación actual.

La mejora propuesta en el modelo tiene el potencial de generar un impacto económico significativo. Basado en análisis preliminares, se estima que una mejora porcentual en el modelo podría haber permitido a la organización ahorrar hasta $1.950.397 en pérdidas asociadas a usuarios no pagadores en un año. Entre agosto de 2023 y febrero de 2024, la utilidad neta registrada con todos los usuarios fue de -$492.273, lo que proyecta una pérdida anual de aproximadamente -$738.409. Por lo tanto, una optimización efectiva del modelo podría no solo mitigar estas pérdidas, sino también transformar una situación negativa en utilidades positivas para la organización.

# 5. Marco conceptual

El desarrollo del presente trabajo se enmarca en el campo de la ciencia de datos, específicamente en el ámbito del aprendizaje de máquinas. En este capítulo, se abordarán los conceptos fundamentales relacionados con estos temas. Primero, se presentará una introducción breve basada en la bibliografía existente. A continuación, se discutirá sobre los modelos utilizados en esta área de trabajo y, finalmente, se ofrecerá una explicación detallada de la metodología empleada en este estudio.

La revolución tecnológica del siglo pasado trajo consigo avances cruciales para la sociedad y las organizaciones actuales. En este contexto, el concepto de Inteligencia Artificial (IA) emergió, descrito por Alan Turing como máquinas capaces de realizar tareas humanas, como se detalla en el artículo de Hadri (2021) titulado *Machine Learning, Data Science and Artificial Intelligence*. Más tarde, se desarrolló el aprendizaje de máquinas o *Machine Learning* (ML), una rama de la IA que se refiere a algoritmos capaces de mejorar su desempeño con la experiencia. Este desarrollo tuvo lugar aproximadamente 20 años después de las ideas propuestas por Turing, como se menciona en el trabajo de Firican (s.f.) titulado *The history of Machine Learning*.

Paralelamente al avance de la IA, surgió el concepto de ciencia de datos o *Data Science* (DS), definido por Mathur (2023) en su artículo *Data Science vs. machine learning: What's the difference?* como el proceso de crear valor a partir de datos. Según Kroese, Botev, Taimre, y Vaisman (2019) en su libro *Data Science and Machine Learning: Mathematical and Statistical Methods*, existe una intersección entre ciencia de datos y aprendizaje de máquinas, dado que la ciencia de datos emplea modelos de aprendizaje de máquinas para analizar datos y entender patrones.

## 5.1. Modelos de aprendizaje de máquinas

Como se ilustra en la Figura 5.1, Wakefield (s.f.) en su artículo *A guide to the types of machine learning algorithms and their applications* clasifica los modelos de ML en cuatro categorías principales: aprendizaje supervisado, semisupervisado, no supervisado y de refuerzo. En el aprendizaje supervisado, el modelo recibe valores de entrada (*inputs*) y salida (*outputs*); en el semisupervisado, puede no recibir todos estos valores; en el no supervisado, no se reciben valores de salida, ya que el objetivo es agrupar datos basados en características similares; y en el de refuerzo, el modelo aprende mediante reglas y múltiples iteraciones. Dado el tipo de datos disponibles en el historial de operaciones, que incluyen características observadas de los usuarios, valores de entrada y comportamiento de pago, se optará por un modelo de aprendizaje supervisado.

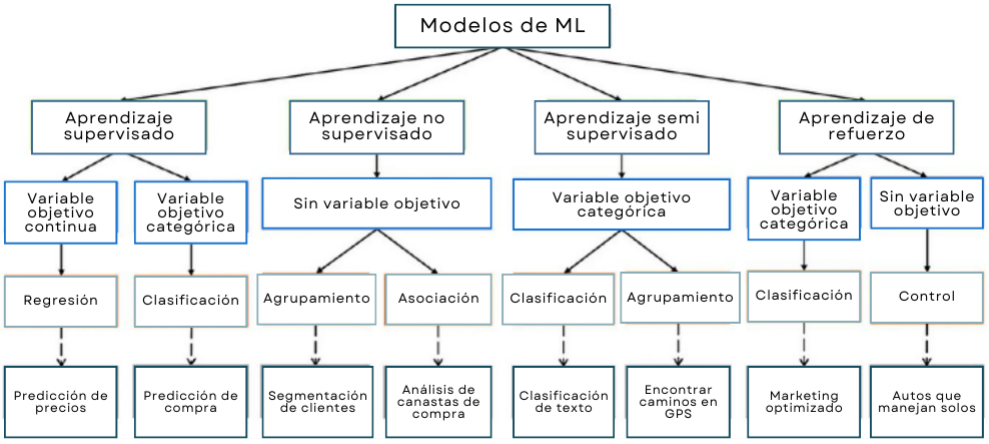


Figura 5.1: Tipos de modelos de ML. Fuente: Adaptado de *List of Machine Learning Models*.- Aamir Kalimi.

Además de seleccionar el tipo de modelo, es necesario decidir entre un modelo de regresión o de clasificación. Según el artículo *Regression & Classification*[22], un modelo de regresión predice una variable continua[23], mientras que un modelo de clasificación asigna datos a un conjunto de valores definidos. En este trabajo, dado que el objetivo es decidir si otorgar o no un crédito, que se puede representar como una asignación de valor de 1 o 0, se utilizará un modelo de **clasificación**.

Dentro de los modelos de clasificación, existe una gran variedad de ellos, como se menciona en el listado de Kalimi (2023). Pérez Rojas (2016) en su memoria *Diseño de metodología para el seguimiento de modelos de riesgo crediticio* destaca la función Logit como un modelo estándar en el análisis de riesgo, representado en la Ecuación 5.1. Esta ecuación permite calcular una probabilidad de pago ($p_i$) utilizando coeficientes($\beta_k$) y variables ($x_{ki}$), mediante la aplicación de una función exponencial, como se muestra en la Ecuación 5.2.

$$\text{Logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} \tag{5.1}$$

$$p(y_i = 1|x_i) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{V} \beta_j x_{ij})}} \tag{5.2}$$

Al la probabilidad calculada se le asigna un umbral, comúnmente 0,5, para clasificar al cliente como bueno (si la probabilidad supera el umbral) o malo (si está por debajo), como menciona Blancas (2022) en su artículo *Stop using 0.5 as the threshold for your binary classifier*. Aunque en el documento se menciona que otros modelos más complejos, denominados *black box*, no permiten una interpretación simple de las variables que influyen en ellos[24], se han desarrollado herramientas como los valores de SHAP para ofrecer una interpretación de las variables [25].

[22] Para más información: https://www.linkedin.com/pulse/regression-classification-dr-saurav-das-dqlmc/
[23] Una variable continua puede tomar infinitos valores en un rango dado (Thomas, 2022).
[24] Un modelo *black box* no proporciona una explicación directa de cómo predice (Gil Martínez, 2020).
[25] Los valores de SHAP se basan en la teoría de juegos cooperativos y se utilizan para determinar la importancia relativa de las variables en un modelo (Trevisan, 2022).

De acuerdo con Alonso y Carbó (2021) y Kryńska (2020) en sus artículos *Understanding the Performance of Machine Learning Models to Predict Credit Default: A Novel Approach for Supervisory Evaluation* y *Comparison of models for credit risk purposes - logistic regression vs random forest*, respectivamente, el modelo Logit es superado en rendimiento por modelos más complejos como Random Forest. De hecho, según Espinosa-Zúñiga (2020) en *Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito*, dos algoritmos destacados para la predicción de riesgo financiero corresponden a bosques aleatorios o *random forests* y XGBoost, este último perteneciente a los algoritmos de potenciación de gradiente o *gradient boosting*.

En la Tabla 5.1, al comparar algunos modelos según el artículo *Balancing Act: The Pros and Cons of Machine Learning Algorithms* de Kumar (2024), se observa que si bien la implementación e interpretación de los algoritmos de bosques aleatorios y de potenciación de gradiente es más compleja en comparación con regresiones y árboles de decisión, estos algoritmos ofrecen un mejor rendimiento al momento de realizar predicciones.

| **Aspectos** | **Regresión lineal y logística** | **Árboles de decisión** | **Bosques aleatorios** | **Potenciación de gradiente** |
|---|---|---|---|---|
| Interpretabilidad | Alta | Alta | Baja | Media |
| Facilidad de implementación | Alta | Alta | Media | Media |
| Flexibilidad | Baja | Media | Alta | Alta |
| Exactitud | Baja | Media | Alta | Alta |
| Tendencia a sobreajuste | No | Sí | No | Sí |
| Otros | Asume linealidad | Sensible a cambios | Alto uso de recursos computacionales | Dependiente de parámetros |

Tabla 5.1: Comparación de modelos de aprendizaje de máquinas. Fuente: *Elaboración propia*

Para proporcionar contexto sobre Random Forest y XGBoost, según Breiman (2001) en su artículo *Random Forests*, este algoritmo es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión. Estos árboles se combinan para obtener un resultado superior al que se obtendría al usarlos de forma individual, como menciona Lizares (2017) en su tesina *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico*. Cada árbol se construye utilizando un subconjunto aleatorio de las observaciones, seleccionadas mediante la técnica de *bootstrap*[26]. Posteriormente, los resultados de los árboles se combinan mediante ensamblado[27], generalmente a través del promedio de los resultados.

Por otro lado, según Chen y Guestrin (2016) en su artículo *Xgboost: A scalable tree boosting system*, el algoritmo XGBoost, o eXtreme Gradient Boosting, también corresponde

[26] La técnica de *bootstrap* se refiere a un método de muestreo de datos con reemplazo realizado varias veces (Rosidi, 2023).

[27] El ensamblado de modelos es una técnica que combina múltiples modelos individuales para mejorar la precisión de las predicciones (Mina Chiquiza, 2024).

a una técnica de aprendizaje supervisado basada en árboles de decisión, pero en este caso, los árboles crecen hasta su máxima extensión. XGBoost es un ensamblaje de árboles de decisión que se genera de forma secuencial y paralela, donde cada árbol creado reduce el error del anterior.

Además de los algoritmos mencionados, LightGBM, o Light Gradient Boosting Machine, tiene un alto potencial para obtener buenos resultados. Según Bastos (2020) y Xu, Ji, Li, y Lü (2023) en sus artículos *Credit Risk Analysis with Machine Learning* y *Small data machine learning in materials science* respectivamente, LightGBM se visualiza como un modelo óptimo en ciertos casos aplicados. LightGBM también se basa en árboles de decisión y es similar a XGBoost; sin embargo, la diferencia entre estos algoritmos, según el artículo *Light GBM vs XGBoost. ¿Cuál es mejor el algoritmo ?* de Barrios Arce (2022), radica en que LightGBM construye estos árboles por nivel, mientras que XGBoost lo hace por hoja, como se muestra en el Anexo C.1.

Así, debido a que los tres modelos mencionados presentan un buen comportamiento en la predicción de clases relacionadas con la entrega de créditos, se decide evaluarlos en esta memoria.

## 5.2. Criterios de evaluación

En el análisis de datos y la modelización predictiva, es esencial contar con criterios específicos tanto para evaluar diferencias dentro de los datos como para medir el rendimiento de los modelos creados. La primera parte de esta sección se enfocará en la evaluación de diferencias entre distribuciones de datos, específicamente sobre el historial de compras de los usuarios, mientras que la segunda parte abordará los criterios de evaluación para comparar el rendimiento de los modelos predictivos a desarrollar.

Al analizar resultados y diferencias entre dos distribuciones, en este caso para resaltar la diferencia entre buenos y malos pagadores, es necesario definir un criterio de evaluación que confirme la relevancia de la diferencia entre ambos grupos. Para ello, según Lorenzo (2019) en su publicación *Estadística básica: Introducción a la Prueba t de Student y el Análisis de la Varianza*, se utiliza la conocida prueba estadística t de Student, que permite comprobar la diferencia entre las medias de dos distribuciones.

También es necesario definir un criterio de evaluación para medir el rendimiento después de la creación de modelos. En la literatura, para la evaluación de modelos de clasificación, existe una gran variedad de métricas. En este sentido, Rainio, Teuho, y Klén (2024) en *Evaluation metrics and statistical tests for machine learning*, Vujović (2021) en *Classification model evaluation metrics*, y Hossin y Sulaiman (2015) en *A review on Evaluation Metrics for Data Classification Evaluations* proponen de forma unánime las métricas basadas en una matriz de confusión.

La matriz de confusión, como se observa en la Tabla 5.3, resume de forma simple las predicciones realizadas por los modelos y las compara con los valores reales, obteniendo así cuatro cuadrantes: verdadero negativo (vn) y verdadero positivo (vp), referentes a las predicciones correctas para la clase negativa y positiva respectivamente, y falso negativo (fn) y falso positivo (fp), que se refieren a predicciones incorrectas donde los valores reales

positivos se asignaron como negativos y los valores reales negativos se asignaron como positivos, respectivamente.

Las métricas basadas en esta matriz incluyen principalmente exactitud, sensibilidad, especificidad, precisión, recuperación y puntuación-F1, cuyas fórmulas se encuentran en la Figura 5.2. Junto con las métricas anteriores, los mismos autores proponen una medida visual referente a la curva *Receiver Operating Characteristic* (ROC). La curva ROC tiene en sus ejes la sensibilidad (también conocida como tasa de verdaderos positivos) y la especificidad (tasa de verdaderos negativos). El área bajo esta curva, conocida como AUC (Area Under the Curve), es un valor que resume la capacidad del modelo para distinguir entre las clases. Un AUC de 1 indica un modelo perfecto, mientras que un AUC de 0.5 indica un modelo que no tiene capacidad de discriminación, equivalente a realizar predicciones al azar.

| **Métricas** | **Fórmula** | **Foco de evaluación** |
|---|---|---|
| Exactitud | $\frac{vp+vn}{vp+fp+tn+fn}$ | Ratio de predicciones correctas sobre el número total de instancias evaluadas |
| Sensibilidad | $\frac{vp}{vp+fn}$ | Fracción de patrones positivos que se clasifican correctamente |
| Especificidad | $\frac{vn}{vn+fp}$ | Fracción de patrones negativos que se clasifican correctamente |
| Precisión | $\frac{vp}{vp+fp}$ | Patrones positivos que se predicen correctamente a partir del total de patrones predichos en una clase positiva |
| Recuperación | $\frac{vp}{vp+fn}$ | Fracción de patrones positivos que se clasifican correctamente |
| Puntuación-F1 | $\frac{vp}{vp+\frac{1}{2}(fp+fn)}$ | Media armónica entre valores de recuperación y precisión |

Tabla 5.2: Tabla de métricas. Fuente: Adaptado de *A review on Evaluation Metrics for Data Classification Evaluations*.- Hossin, M. y Sulaiman, M.

| | | Clase predicha | |
|---|---|---|---|
| | | Negativa | Positiva |
| Clase real | Negativa | Verdadero negativo | Falso positivo |
| | Positiva | Falso negativo | Verdadero positivo |

Tabla 5.3: Matriz de confusión. Fuente: *Elaboración propia*

Si bien sería correcta la utilización de cualquiera de las métricas mencionadas, se optará por:

- **Valores AUC:** Esta métrica se elige por su interpretabilidad y su capacidad para incorporar la minimización de errores de clasificación tipo I y II[28].
- **Tasa de dinero prestado a falsos positivos:** Esta métrica se refiere al porcentaje de dinero prestado a malos pagadores sobre el total prestado. Se utiliza como principal

[28] Según Velayudhan (2020), existe ambigüedad en el uso de estos términos; no obstante, se tomará como error de tipo I a los falsos positivos y de tipo II a los falsos negativos.

métrica por razones económicas y con visión de negocio, dado que, según Abdou y Pointon (2011) en su artículo *Credit scoring, statistical techniques and evaluation criteria: A review of the literature*, los errores al clasificar incorrectamente a una persona con altas probabilidades de no pagar tienen un impacto entre 5 a 10 veces mayor que negar un crédito a una persona con altas probabilidades de pago.

- **Utilidad:** Esta medida de evaluación se utiliza con el objetivo de aumentar las ganancias del servicio BNPL para Cleo. El cálculo de la utilidad corresponde a la diferencia entre los ingresos generados por cargos de servicio a los usuarios y los costos de obtener la información de la persona y las pérdidas por realizar una mala predicción.

## 5.3. Software y librerías

El software utilizado por la organización es Python, un lenguaje de programación de código abierto ampliamente utilizado a nivel global. Algunas de las ventajas de este lenguaje, según menciona Visus (2020) en su artículo *¿Para qué sirve Python? Razones para utilizar este lenguaje de programación*, son que es gratuito, flexible y cuenta con una gran base de colaboradores. Muchos de estos colaboradores contribuyen al desarrollo de su comunidad a través de la creación de librerías, las cuales facilitan el desarrollo del trabajo con este programa.

Al año 2024, existen librerías que contribuyen a facilitar el desarrollo de modelos de predicción y su interpretación. Algunas de estas son Feature-engine, Optuna y SHAP, las que se describen de forma breve a continuación.

- **Feature-engine:** Según Galli (2020) en *Feature-engine: A new open source Python package for feature engineering*, esta herramienta incorpora una gran cantidad de técnicas utilizadas en la ingeniería de características y que pueden ser utilizadas de forma sencilla.

- **Optuna:** Según Akiba, Sano, Yanase, Ohta, y Koyama (2019) en *Optuna: A next-generation hyperparameter optimization framework*, Optuna es un software de optimización de hiperparámetros de modelos de aprendizaje de máquinas; se basa en la maximización o minimización de una función objetivo dada por el usuario.

- **SHAP:** Como menciona Trevisan (2022) en *Using SHAP values to explain how your machine learning model works*, SHAP se utiliza para aumentar la transparencia y la interpretabilidad de los modelos de aprendizaje automático.

## 5.4. Comparación de metodologías

Para la elección del marco de trabajo a utilizar, se analiza una investigación realizada por Saltz (2022) a más de 100 empresas sobre las estructuras metodológicas más utilizadas en proyectos de ciencia de datos. En este estudio, se observa que los métodos más utilizados son CRISP-DM, Scrum y Kanban, por lo que se procede a compararlos en la Tabla 5.4, además de realizar una breve descripción de cada una a continuación.

- **CRISP-DM:** Según Talaviya (2023) en su artículo *CRISP-DM framework: A foundational data mining process model*, esta metodología fue creada en 1996 por

colaboradores de DaimlerChrysler, SPSS y NCR. CRISP-DM es una metodología estándar para el análisis de datos y la minería de datos. Consiste en seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

- **Scrum:** Según Talreja (2023) en *Scrum Origins and Agile Principles: The Foundations of Agile Project Management*, Scrum fue formulado en 1995 por Jeff Sutherland y Ken Schwaber. Es una metodología ágil utilizada principalmente en el desarrollo de software. Se basa en ciclos de trabajo cortos llamados sprints, que generalmente duran entre 1 y 4 semanas. Cada sprint tiene objetivos específicos y se evalúa su desempeño al finalizar. Scrum requiere una actualización constante, con reuniones diarias para discutir el progreso y los impedimentos.

- **Kanban:** Schwartz (2024) en su artículo *Kanban History: Origin & Expansion Across Industries* menciona que Kanban fue concebido en la década de 1940 por Taichi Ohno. Es una metodología visual para la gestión de proyectos que utiliza tarjetas en un tablero para representar las tareas y su estado (pendiente, en progreso, completado). Se enfoca en la mejora continua mediante la visualización del flujo de trabajo y la identificación de cuellos de botella.

| **Aspectos** | **CRISP-DM** | **Scrum** | **Kanban** |
|---|---|---|---|
| Flexibilidad | Alta, cada fase tiene sus propios objetivos y se adaptan a medida que avanza el proyecto | Media, cada ciclo tiene sus objetivos y al terminar se evalúa su desempeño | Baja, solo cambian los estados al estar en proceso o terminada una actividad |
| Demora de implementación | Baja | Alta | Baja |
| Requiere actualización constante | No necesariamente | Sí, diaria | Sí |
| Documentación | Detallada | Variable | Variable |
| Iteración | Sí, en todas las fases | Sí, al completar ciclos | No |
| Enfoque en el negocio | Alto | Medio | Bajo |

Tabla 5.4: Comparación de metodologías. Fuente: *Elaboración propia*

Considerando las ventajas y desventajas de estas tres metodologías, se opta por utilizar la metodología CRISP-DM. Esta decisión se basa en su capacidad de implementación directa en el trabajo y su facilidad de adaptación a las limitaciones temporales, lo que la hace ideal para un proyecto de ciencia de datos. En la Sección 6 se detalla esta metodología con mayor profundidad. Además, según autores como Minoli (2020) en su artículo *Adaptación de Scrum para el desarrollo de soluciones de Business Analytics con CRISP-DM*, CRISP-DM puede combinarse con elementos de Scrum, permitiendo aprovechar los beneficios de ambas estructuras de trabajo.

# 6. Metodología

Para el desarrollo del proyecto, se empleará la metodología CRISP-DM, mencionada al final de la Sección 5.4. Esta metodología se basa en seis grandes pasos, como se ilustra en la Figura 6.1. Según menciona Peralta (2014) en su artículo de la Revista Latinoamericana de Ingeniería de Software, este modelo reconoce la naturaleza cíclica del proceso de explotación de datos y no establece secuencias rígidas entre las fases.

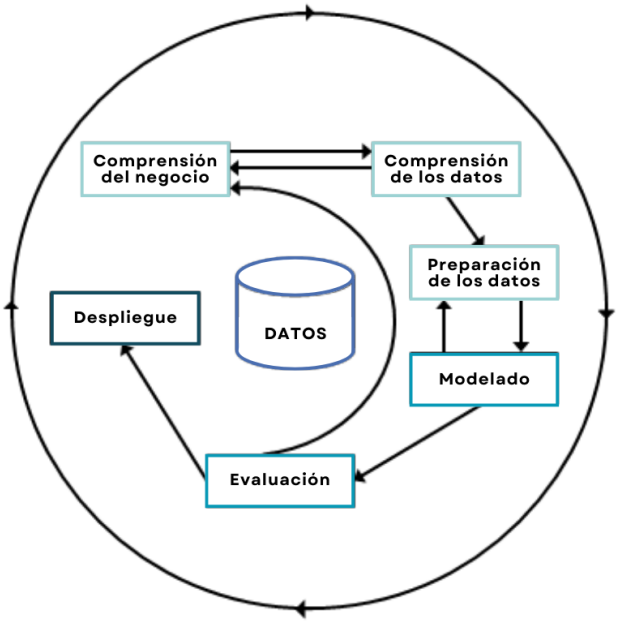


Figura 6.1: Diagrama CRISP-DM. Fuente: Adaptado de *Proceso de conceptualización del entendimiento del negocio para proyectos de explotación de información.*- Federico Carlos Peralta.

A continuación, se detalla cómo se abordará cada una de estas fases en el proyecto, recordando que el proceso no necesariamente se realiza de forma lineal, por lo que hay pasos que pueden repetirse en el ciclo.

- **Comprensión del negocio:** Esta etapa implica actividades como la presentación e inducción del estudiante al proyecto por parte de la organización y la comprensión de la industria Fintec. También se incluye la definición del proyecto y las necesidades de la empresa, además de reuniones con el equipo de BNPL e investigación sobre la compañía.

- **Comprensión de los datos:** Esta etapa se enfoca en entender los datos disponibles, incluyendo el estado del modelo previo y sus resultados, así como el tipo de datos que la compañía pueda conseguir para realizar el proyecto. La organización dispone del modelo "V4", utilizado al inicio del desarrollo de esta memoria, así como de datos existentes de las órdenes realizadas hasta la fecha. El modelo "V4" servirá como punto de partida para la creación del nuevo modelo, utilizando sus predicciones e indicadores como referencia mínima para el rendimiento deseado en el proyecto. Los datos existentes constituirán la base para evaluar cualquier iteración en la fase de modelado, asumiendo que los nuevos

consumidores tendrán un comportamiento similar a los que ya han utilizado el servicio BNPL.

- **Preparación de los datos:** Esta etapa, que suele ser una de las más extensas en proyectos de ciencia de datos, abarca desde la limpieza de los datos hasta el proceso de ingeniería de características. También incluye la investigación sobre modelado y predicción. Particularmente, consta de la selección y limpieza de datos, así como la creación y transformación de variables.

- **Modelado:** En esta fase, se seleccionan uno o varios modelos a utilizar en cada iteración, y utilizando los datos filtrados y limpios, se realizan predicciones sobre el resultado deseado. Incluye el ajuste de hiperparámetros de los modelos. Para esta memoria en específico, el modelado se realizará en tres iteraciones con el fin de obtener el mejor modelo basado en múltiples combinaciones de algoritmos, variables y parámetros, proceso que se muestra de forma gráfica en la Figura 6.2.

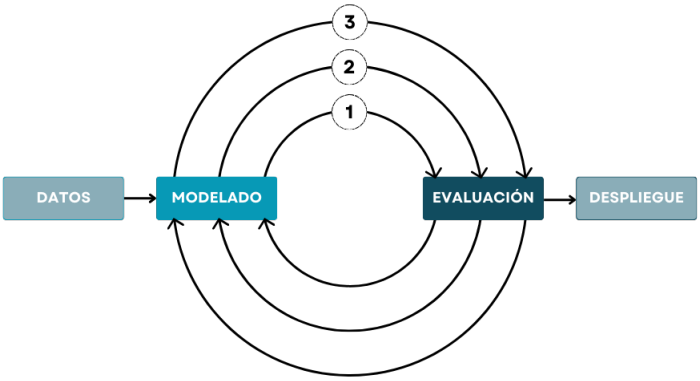


Figura 6.2: Representación de iteraciones. Fuente: *Elaboración propia*

- **Evaluación:** Los modelos de cada una de las iteraciones, como se observa en la Figura 6.2, se evaluarán en función de las métricas presentadas en la Sección 5.2, referentes a la utilidad, la tasa de falsos positivos (FP) y los valores AUC. Además, se llevará a cabo una evaluación y validación cualitativa junto con los miembros del equipo de BNPL.

- **Despliegue:** Debido a los alcances y las fechas del proyecto, no es posible completar la fase de despliegue, que correspondería a implementar el modelo en producción. No obstante, para este proyecto, se considera esta fase como la preparación del modelo para producción. El culmen de esta etapa sería el empaquetado del modelo, junto con el traspaso de los archivos y la documentación correspondiente.

# 7. Desarrollo y resultados

## 7.1. Comprensión del negocio

En esta primera etapa, se dio importancia a entender la situación actual de Cleo. Para comprender su funcionamiento, se realizaron reuniones con la CCO y con los *product owners* de los servicios de la empresa, además de equipos de otras áreas. Junto a esto, se llevó a cabo el análisis del apartado de antecedentes referidos en la Sección 1, del cual se pudo concluir que, si bien la organización se encuentra en una posición favorable con respecto a variables macroeconómicas, está siendo sobrepasada por sus competidores.

Opciones para explotar y volver a su posición de liderazgo incluyen aprovechar nuevas tendencias como la compra en físico, mejoras en la seguridad y en la experiencia del cliente. Al mismo tiempo, es crucial abordar factores que ponen en riesgo su posición, como la dependencia del talento, que requiere estrategias de retención; la falta de recursos, que necesita una estrategia de inversión; y el cambio constante de las regulaciones, para lo que se requiere una mayor flexibilidad en las operaciones.

En las reuniones se entablaron conversaciones sobre el rendimiento del modelo actual, haciendo un énfasis en la baja de este y recalcando que los usuarios que pagan a tiempo no estaban generando ganancias, solo los que pagan atrasados. Así, se llegó a la conclusión de que el nuevo modelo debe reducir la tasa de no pago a menos de un 8 % para generar ganancias con todo tipo de usuarios, analizados en la próxima sección. También se concluyó que, para lograr un modelo con una menor cantidad de no pagos, es probable que sea más restrictivo que el modelo actual en términos del número de personas aceptadas.

## 7.2. Comprensión de los datos

Los datos que obtiene Cleo consisten básicamente en transaccionales, o de la orden, sociodemográficos del usuario, y sobre la situación financiera de la persona, según registros de la CMF y de su banco. Estos datos se muestran listados en la Figura 7.1. Para el análisis, se utilizó una base con datos en Excel desde octubre de 2019 hasta febrero de 2024.

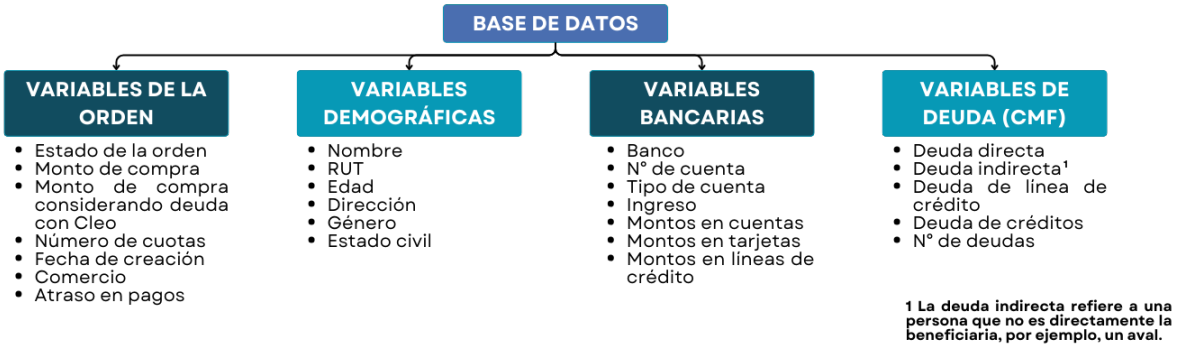


Figura 7.1: Datos con mayor relevancia. Fuente: *Elaboración propia*

Esta fase se centró principalmente en conocer el comportamiento de los usuarios de Cleo,

analizando las órdenes realizadas, con el fin de determinar si existe alguna diferencia significativa en las variables que pudiera permitir segmentar a los tipos de usuarios.

Así, el análisis realizado en esta sección se basó principalmente en la edad del usuario, el monto de la compra y su estado financiero, en términos de ingresos, saldo y deudas. Para llevar a cabo esta exploración de los datos de usuarios, se analizaron las tendencias mensuales entre los dos tipos de usuarios, buenos y malos, entre julio de 2022 y enero de 2024, resultando en las gráficas representadas en el Anexo D.1. Además de analizar de forma gráfica los datos, se realizaron pruebas t de Student con el objetivo de concluir si existía alguna diferencia estadística entre ambos grupos de personas.

Los resultados, presentes en la Tabla 7.1, fueron concluyentes y mostraron que solamente la edad no tiene una diferencia significativa entre los buenos y malos usuarios, mientras que el monto de compra, la deuda, el ingreso mensual y el saldo en la cuenta, en la tarjeta y en la línea de crédito sí presentaban diferencias significativas entre ambos grupos según la prueba t de Student con un valor de corte de 0.05[29]. Así, contando con variables que tienen una diferencia suficiente para separar los grupos, se emplearon en la posterior sección.

| **Variable** | **P-valor** | **Significancia** |
|---|---|---|
| Edad | $0,447$ | No hay evidencia suficiente |
| Monto de compra | $0,009$ | Hay diferencia significativa |
| Deuda directa | $0,020$ | Hay diferencia significativa |
| Deuda de línea de crédito | $0,005$ | Hay diferencia significativa |
| Ingreso mensual | $0,024$ | Hay diferencia significativa |
| Saldo en cuenta | $0,002$ | Hay diferencia significativa |
| Monto total: tarjeta | $2,573{\cdot}10^{-9}$ | Hay diferencia significativa |
| Monto disponible: tarjeta | $3,643{\cdot}10^{-5}$ | Hay diferencia significativa |
| Monto total: línea de crédito | $7,799{\cdot}10^{-8}$ | Hay diferencia significativa |
| Monto disponible: línea de crédito | $1,558{\cdot}10^{-4}$ | Hay diferencia significativa |

Tabla 7.1: Resumen diferencias entre buenos y malos usuarios. Fuente: *Elaboración propia*

## 7.3. Preparación de los datos

Para preparar los datos de forma adecuada para el modelo, se siguió el flujo de trabajo representado en la Figura 7.2. La base de datos inicial entregada por la compañía consta de un archivo CSV de 203 variables con 8,463 registros, entre los que se incluyen datos personales del usuario, bancarios, de la CMF y de la compra en sí; los datos del usuario, bancarios y de la CMF son obtenidos a través de una empresa que utiliza los protocolos de *Open Banking*[30].

[29] La aceptación de la hipótesis de que existe una diferencia entre las medias de los grupos se basa en que el P-valor sea menor que el escogido para realizar la prueba.

[30] Término referido a compartir los datos financieros entre varias instituciones con el objetivo de mejorar y personalizar ofertas. Esta práctica se basa en el consentimiento del cliente (Maldonado Rosas, 2021).

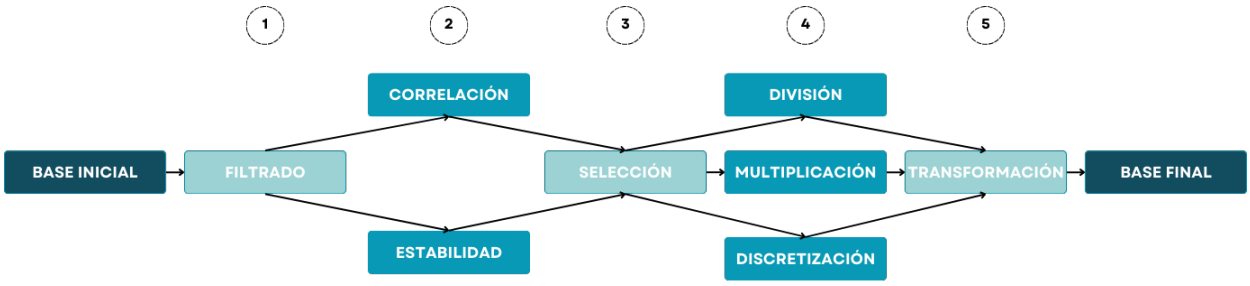


Figura 7.2: Flujo de preparación de datos. Fuente: *Elaboración propia*

En esta gráfica, se presentan cinco pasos principales, detallados en el Anexo D.2, que de forma resumida serían:

1. En el primer paso, se filtró la base por las compras confirmadas, eliminando las que estuvieran en un período de pago vigente, ya que para estas compras no se puede discernir sobre su estado final. Así, resulta una base de datos de 6,168 observaciones.

2. En el segundo paso, se utilizaron herramientas de la librería Feature-engine para elegir variables según su correlación y su índice de estabilidad. Estas librerías permiten buscar las variables con mayor valor para el modelo y que no modifiquen de forma súbita los resultados, lo que era buscado para conseguir valores consistentes.

3. En el tercer paso, se realizó el cruce de ambos métodos de selección de variables para escoger una base coherente y en pos de evitar el posterior sobreajuste de modelos. También se eliminaron las filas sin datos o con valores nulos, y no resultó necesario cambiar el tipo de las variables, ya que la totalidad de las utilizadas fueron del tipo numérico. Se seleccionaron 26 variables y 2,951 registros, detalladas en la Tabla 7.2.

| **Tipo** | **Variables** | **Rangos/valores** |
|---|---|---|
| CMF | • Deuda directa<br>• Deuda indirecta<br>• Deuda de créditos | • $[0, 411.793.469]$<br>• $[0, 114.785.866]$<br>• $[0, 97.730.398]$ |
| Ingreso y saldo | • Ingreso mensual<br>• Monto de dinero en cuenta<br>• Límite de dinero en tarjetas<br>• Límite de dinero en líneas de crédito | • $[0, 36.693.333]$<br>• $[0, 80.800.201]$<br>• $[0, 30.000.000]$<br>• $[0, 15.000.000]$ |
| Compra y personales | • Monto de compra<br>• Número de cuotas<br>• Edad | • $[0, 991.389]$<br>• $\{1, 3, 6\}$<br>• $[18, 77]$ |

Tabla 7.2: Resumen de variables a utilizar. Fuente: *Elaboración propia*

4. En el cuarto paso, se ampliaron las variables seleccionadas mediante criterios de división, multiplicación y discretización. Estos procesos son útiles para reducir el impacto de errores y capturar relaciones no lineales entre variables (Mondal, 2024). Al finalizar este paso, la base de datos volvió a contar con 203 variables.

5. En el quinto paso, se transformaron las variables resultantes a una distribución normal y se estandarizaron para que tuvieran una magnitud similar. Esta transformación ayuda a minimizar el impacto de valores atípicos y a mejorar la interpretación (Urrego, 2023).

Con estos pasos, se concluyó la fase de preparación con una base de datos ajustada, con datos menos correlacionados, variables que capturan una mayor información a través de interacciones, y valores estandarizados, lo que permitió reducir posibles sesgos en el modelado.

## 7.4. Modelado

Como se mencionó previamente en la Sección 5.1 del Marco Conceptual, los modelos diseñados en esta sección incluyen Random Forest, XGBoost y LightGBM. Estos algoritmos se seleccionaron por su rendimiento y su amplia utilización en proyectos similares. Random Forest fue elegido por su robustez y capacidad para manejar grandes volúmenes de datos, aunque puede ser costoso en términos computacionales. XGBoost y LightGBM se seleccionaron por su alto rendimiento y control de ajuste, aunque requieren una cuidadosa sintonización de parámetros. Además, LightGBM tiene una menor cantidad de documentación en comparación con los otros dos, dado que es más reciente.

Al utilizar una metodología iterativa, esta fase incluye varias etapas, presentadas visualmente en la Figura 7.3 y listadas a continuación en el orden de realización de forma resumida, también se listan en el Anexo D.3 de manera extendida.

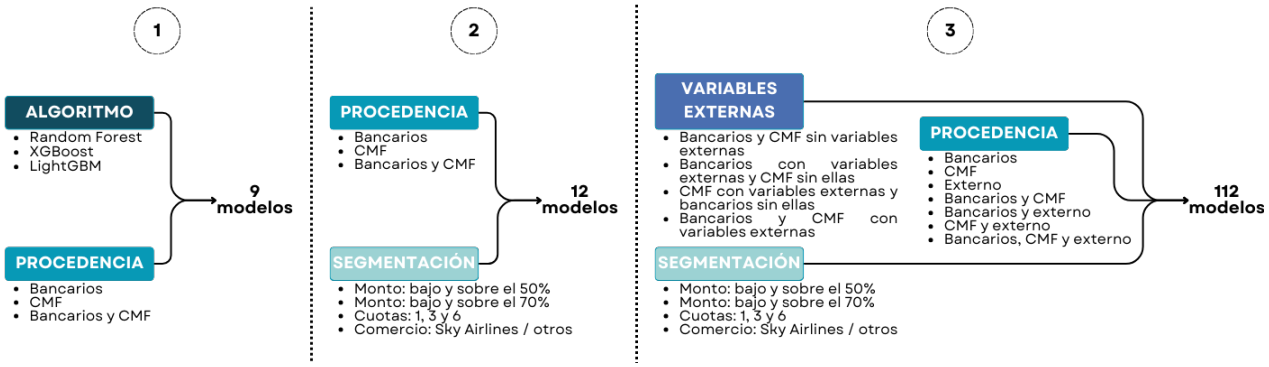


Figura 7.3: Componentes del modelado. Fuente: *Elaboración propia*

- En la primera iteración, se usaron los algoritmos Random Forest, XGBoost y LightGBM con la base de datos completa. Para cada uno de ellos, se diseñaron tres modelos basados en la procedencia de los datos: uno con datos bancarios e ingresos, otro con datos de deuda en la CMF, y un tercer modelo que corresponde al ensamblaje de estos dos.

- En la segunda iteración, se seleccionó el mejor algoritmo del paso previo, y basándose en este, se diseñaron nuevos modelos considerando la segmentación de la base de datos y la separación previa de la procedencia de los datos. Luego, se compararon los resultados de esta fase con los del modelo actual. Las segmentaciones incluyeron usuarios con montos de compra bajo y sobre del percentil 50 y 70, que compran en 1, 3 y 6 cuotas, y una segmentación específica para uno de los clientes más relevantes de Cleo, Sky Airlines.

- En la tercera iteración, se utilizaron las mismas segmentaciones que en la segunda iteración, pero se añadieron variables externas de una empresa de seguridad,

correspondientes a puntajes basados en el riesgo de fraude, obtenidos a partir de patrones de navegación y comportamiento en internet.

Así, se obtuvieron tres sets de modelos para su posterior evaluación. Cabe destacar que cada modelo construido utiliza la librería Optuna para optimizar y mejorar sus predicciones.

## 7.5. Evaluación

En la etapa de evaluación, se emplearon los criterios mencionados en la Sección 5.2 para cada uno de los modelos generados en cada iteración, considerando los valores de AUC, la tasa de dinero entregado a falsos positivos y la utilidad potencial del modelo si se hubiera implementado desde el inicio de las operaciones. Por razones de negocio y de interés para la empresa, la utilidad y la tasa de falsos positivos fueron los indicadores principales. Además de la evaluación cuantitativa basada en métricas, se evaluó cualitativamente si los modelos presentaban algún sesgo basado en las variables utilizadas para su creación.

### 7.5.1. Fase 1: Elección de algoritmo

Luego de poner a prueba los distintos modelos para cada algoritmo, descritos en la Sección 7.4 y con los resultados obtenidos (Detalles disponibles en Anexo D.1), se pudo concluir que, aunque los modelos basados en Random Forest y LightGBM ofrecieron una mayor utilidad que los basados en XGBoost, este último presentó una tasa de no pago significativamente menor, entre 5 y 30 veces dependiendo del modelo. En términos de AUC, los modelos individuales de Random Forest y LightGBM superaron a XGBoost, mientras que en el ensamble el rendimiento fue el opuesto.

En general, los modelos de ensamble mejoraron el rendimiento de los algoritmos, tal como se tenía previsto según la literatura. Al ser esto así, y comparando los ensambles de cada algoritmo, con sus resultados presentes en la Tabla 7.3, se seleccionó el modelo basado en XGBoost debido a su mejor rendimiento en métricas clave. Aunque este mostró una utilidad aproximadamente un 3 % menor que los otros modelos, sus resultados en AUC y en la tasa de no pago fueron superiores, siendo esta última cinco veces menor en el modelo con XGBoost en comparación con los competidores.

| **Algoritmo** | **Utilidad** | **No pago** | **AUC** |
|---|---|---|---|
| Random Forest | $7.750.523 | 0,696 % | 0,954 |
| XGBoost | $7.574.309 | 0,141 % | 0,969 |
| LightGBM | $7.801.970 | 0,677 % | 0,957 |

Tabla 7.3: Comparación resultados modelos de ensamble Random Forest, XGBoost y LightGBM. Fuente: *Elaboración propia*

Las matrices de confusión de este modelo XGBoost con ensamble se encuentran en la Figura 7.4, donde se observa que las predicciones evitan los falsos positivos en comparación con los falsos negativos, lo cual es favorable en términos de negocio.

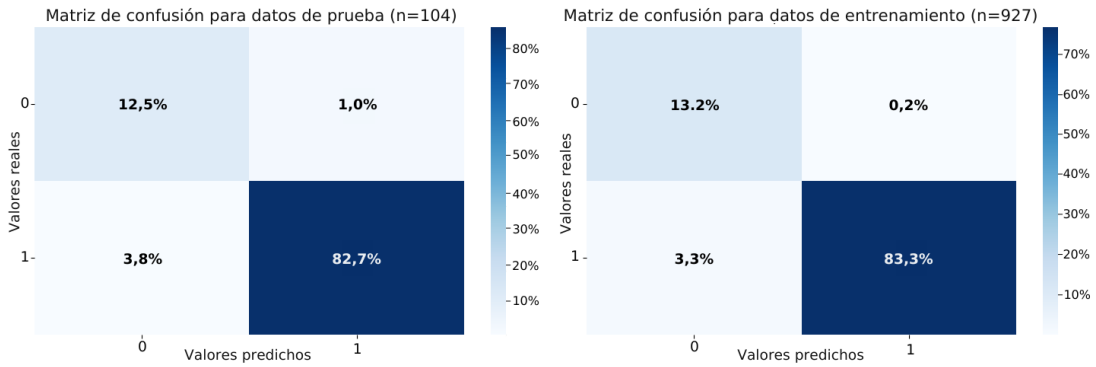


Figura 7.4: Matrices de confusión del modelo escogido: XGBoost, Bancario + CMF. Fuente: *Elaboración propia*

### 7.5.2. Fase 2: Evaluación con respecto a datos reales

Una vez seleccionado el algoritmo a utilizar, XGBoost, se compararon los resultados de las distintas segmentaciones y modelos con los resultados reales obtenidos por la compañía entre agosto de 2023 y febrero de 2024, cuando se utilizó el modelo actual "V4". Dado que esta iteración corresponde a la comparación con los resultados del modelo actual y su rendimiento, tema central de esta memoria, se listan a continuación las comparaciones de las métricas reales con las obtenidas en esta fase. Para estas comparaciones se tomaron en cuenta los resultados detallados, presentes en el Anexo D.2, y los de los mejores modelos, presentados en la Tabla 7.4. En esta tabla, los modelos se ordenan por utilidad, siendo el primero el de mayor utilidad y AUC, mientras que el tercero corresponde al de menor tasa de no pago.

| **Modelo** | **Segmentación** | **Utilidad** | **No pago** | **AUC** |
|---|---|---|---|---|
| Bancario + CMF | 50 / 50 | $723.726 | 0,741 % | 0,962 |
| Bancario + CMF | Sky / No Sky | $686.842 | 1,058 % | 0,941 |
| Bancario | 50 / 50 | $666.622 | 0,000 % | 0,929 |

Tabla 7.4: Mejores modelos para comparar con datos reales. Fuente: *Elaboración propia*

- **Tasa de dinero prestada a falsos positivos:** Según los datos presentados en la Figura 2.1, la tasa de no pago que mostró el modelo para los meses mencionados fue de 8,476 %. Comparando esta información con los resultados de esta iteración, que rondan entre 0 % y 4,268 %, se observó una mejora significativa, con una reducción de más del 50 %. Además, en comparación con los modelos de la Tabla 7.4, la mejora supera el 80 %.

- **Utilidad:** Según los valores mostrados al final de la Sección 4 de alcances, la utilidad generada por el modelo actual en el intervalo de meses mencionado fue de -$492.273. En este sentido, dado que todos los modelos obtenidos presentaron valores positivos, se observó una mejora en esta métrica, con valores que van desde $118.555 hasta $723.726, siendo este último el más alto, como se muestra en la Tabla 7.4.

- **AUC:** El valor AUC del modelo actual se obtuvo considerando a las personas aceptadas en comparación con aquellas que resultaron ser buenas o malas pagadoras, resultando en 0,5. En comparación con los resultados de los modelos, que varían entre 0,667 y 0,962, se observó una mejora considerable de más del 90 % respecto al mejor resultado.

Así, como se ha comprobado de forma detallada para cada métrica, los modelos obtenidos presentaron mejores resultados que el modelo actual, por lo que se puede considerar que el objetivo principal de esta memoria ha sido alcanzado.

Para seleccionar el mejor modelo en esta fase, se observó nuevamente una superioridad en los de ensamble, destacándose la segmentación 50/50 por su rendimiento. Por lo tanto, y considerando razones de negocio, se propone el modelo con la segmentación que ofrece la mayor utilidad y valor AUC, además de una tasa de no pago inferior al 1 %. Este modelo corresponde al ensamble de datos bancarios y de la CMF, segmentado a la mitad por el monto de compra. Las matrices de confusión de este modelo se presentan en la Figura 7.5, lo que permite visualizar la asignación de valores en los datos de prueba y entrenamiento

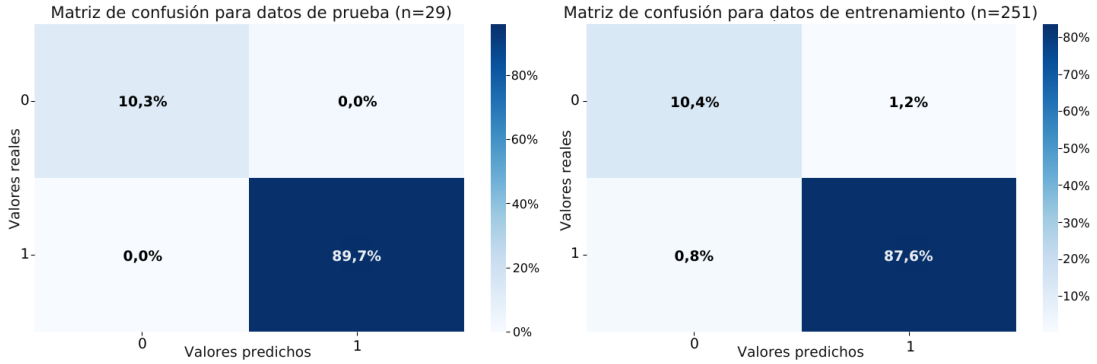


Figura 7.5: Matrices de confusión del modelo escogido: Segmentación 50/50, Bancario + CMF. Fuente: *Elaboración propia*

### 7.5.3. Fase 3: Integración de variables externas

En esta última iteración, se intentó mejorar el resultado anterior, añadiendo variables externas proporcionadas por una empresa de seguridad a los modelos ya utilizados, y creando un nuevo modelo exclusivamente con estas variables, utilizado de manera individual y en ensamble con los otros modelos. Debido a la extensión de los resultados, estos se muestran en los Anexos D.3, D.4 y D.5.

Se obtuvieron 92 modelos que incluían las variables externas y 20 combinaciones sin ellas. Dada la gran diferencia entre ambos conjuntos, se seleccionaron los 10 mejores modelos de cada conjunto para comparar el promedio de sus métricas, excluyendo así los modelos con peor desempeño, ya que no aportan valor significativo a la comparación.

La comparación del promedio de los modelos con mejor rendimiento se presenta en la Tabla 7.5, donde se observa que, en general, los que incorporan las variables de la empresa de seguridad muestran una mejora aproximada del 35 % tanto en utilidad como en la tasa de no pago, en comparación con aquellos que utilizaron solo las variables de Cleo. Aunque el valor AUC también mejora, el incremento es de solo un 5 %. Es importante tener en cuenta que, aunque las variables externas añaden un costo fijo por transacción de aproximadamente un 0,4 % del valor de la UF más IVA, el resultado final sigue siendo favorable. Un análisis más detallado se encuentra en el Anexo D.6.

| Uso de variables externas | Utilidad | No pago | AUC |
|---|---|---|---|
| Sí | $ 7.648.111 | 0,479 % | 0,969 |
| No | $ 5.468.469 | 0,729 % | 0,914 |

Tabla 7.5: Cotejo de uso de variables externas. Fuente: *Elaboración propia*

De los mejores modelos, el que utiliza un triple ensamble, además de incluir la segmentación por comercio, se propone para su futura implementación. Este modelo presentó la mayor utilidad y uno de los AUC más altos, de 0.973, con una tasa de no pago a falsos positivos alrededor del 0.5 %, solo superada por otros modelos con menor utilidad y AUC. Las matrices de confusión de este modelo se muestran en la Figura 7.6, donde se aprecia un error alrededor del 2 % en la asignación de valores para los conjuntos de prueba y entrenamiento.

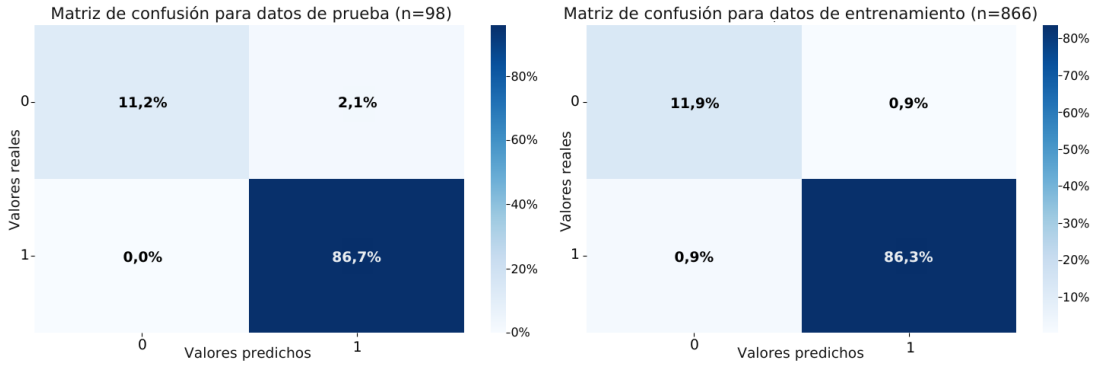


Figura 7.6: Matrices de confusión del modelo escogido: Segmentación Sky / No Sky y triple ensamble. Fuente: *Elaboración propia*

### 7.5.4. Evaluación de sesgo

Para esta parte, se consideró la herramienta para la medición de sesgos y equidad desarrollada por el Laboratorio de Gobierno de la Universidad Adolfo Ibáñez[31]. Esta herramienta se basa en la utilización de variables protegidas conforme a la ley de no discriminación, como raza y sexo, entre otras. Dado que ninguna de las variables utilizadas en los modelos de Cleo entra en esta categoría, se concluyó que no existe sesgo directo en la ejecución, por lo que se seleccionó el último modelo presentado como el óptimo.

Esta fase cobra gran importancia en la actualidad, ya que las variables protegidas son ampliamente consideradas en las empresas. En Cleo, se prioriza la ausencia de estas variables, lo que añade un valor social a su propuesta. Sin embargo, la variable de ingreso podría potencialmente introducir un sesgo indirecto relacionado con el sexo, por ejemplo. A pesar de esto, dicha variable es fundamental en el modelamiento de créditos y no puede ser eliminada. Por lo tanto, se implementarán monitoreos periódicos para mitigar cualquier sesgo inherente.

## 7.6. Despliegue y evaluación económica

En esta última etapa de la metodología, se procedió al empaquetado del modelo, guardando las variables seleccionadas, las transformaciones aplicadas a la base de datos y los modelos utilizados en archivos individuales. Posteriormente, se modificó el archivo de

[31] Para más información: https://herramienta-sesgos-equidad-goblab-uai.streamlit.app/

ejecución del modelo actual en Python para incorporar el nuevo modelo y los archivos del paquete, asegurando la coherencia en el formato de los modelos. Finalmente, el modelo fue entregado a la empresa, ofreciendo al estudiante la posibilidad de continuar su trabajo en la organización de forma indefinida una vez completado el proceso de titulación.

Además, se propusieron los pasos a seguir en la implementación (Detalles en Anexo D.7), destacando la necesidad de establecer un sistema de monitoreo periódico y actualización de datos cada 3 meses, alineado con la frecuencia de renovación de datos actual. Este sistema requeriría una semana de trabajo mensual de un colaborador con conocimientos en ciencia de datos para el monitoreo, lo que equivaldría a una cuarta parte de su sueldo, estimado en aproximadamente $400.000 según Chiletrabajos[32]. Además, se necesitaría otra semana adicional al tercer mes para la actualización de los datos.

Al finalizar el proyecto, se realizó un análisis costo-beneficio. Por un lado, el costo agregado corresponde a la inclusión de las variables externas, lo que tendría un valor de 0,04 % de una UF más IVA, equivalente a $177 al momento del cálculo. Este valor fue entregado al estudiante por parte del equipo de BNPL, considerando negociaciones previas. Además, considerando los costos del nuevo encargado, se agregarían $533.333 mensuales.

Por otro lado, el principal beneficio del modelo es la reducción de la tasa de no pago, de un 8,5 % a un 0,7 %, según el mejor de la Tabla D.3. Usando un promedio de órdenes de $85.080 y un promedio de 2.658 órdenes mensuales según los datos, junto con las comisiones de los comercios (3 %) y usuarios (6 %), el cálculo costo-beneficio correspondería a:

$$\{[(8,5\,\% - 0,7\,\%) \cdot \$85.080 \cdot 9\,\%] - \$177\} \cdot 2.658 - \$533.333 = \$583.027 \tag{7.1}$$

La diferencia calculada es positiva, lo que sugiere que la implementación del nuevo modelo resultará en mayores ganancias. Además, se presume que si se produjera un aumento en el número de órdenes por una expansión de Cleo o cambios en el comportamiento de los usuarios, *ceteris paribus*, la utilidad calculada podría incrementar aún más.

## 7.7. Resumen del capítulo

Las fases de la metodología, de forma directa o indirecta, se relacionaron con la consecución de los objetivos. Las fases de comprensión se vincularon con el análisis de los datos de los usuarios; las de preparación y modelamiento, con el diseño de modelos; y las de evaluación y despliegue, con la preparación y validación. La revisión de la literatura, por su parte, se abordó a lo largo de las seis fases que componen CRISP-DM.

Algunos de los hallazgos más importantes incluyen la definición de un límite para la generación de utilidades en relación con la tasa de no pago, la evaluación de las variables según su diferencia entre tipos de usuarios, el cambio de rendimiento entre los modelos simples y los de ensamble, y la mejora que representa el uso de las variables externas.

De forma particular, se presenta un mayor detalle en el Anexo D.8, el cual detalla la completitud de los objetivos alcanzados a través de las fases de la metodología, junto con los hallazgos más importantes de cada una de estas fases.

[32] Para más información: https://www.chiletrabajos.cl/sueldos/data/scientist

# 8. Discusión

En esta sección se analizan las implicaciones de las decisiones tomadas durante el trabajo, con énfasis en la solución escogida, el cumplimiento de los objetivos, las limitaciones, la metodología y los resultados obtenidos.

## 8.1. Solución

La solución seleccionada para abordar la tasa de no pago se centró en la identificación precisa de buenos pagadores, lo cual impactó positivamente. Sin embargo, no se consideraron aspectos estratégicos como la focalización en comercios con menor tasa de no pago, la reestructuración del servicio para incluir más autenticaciones de seguridad, mayores restricciones de compra, o estrategias que permitan a los usuarios evaluar mejor su deuda con Cleo, similar a instituciones financieras.

Cualquiera de las soluciones alternativas mencionadas tendría la ventaja de abordar el problema de un modo más amplio y flexible, pero también requerirían un mayor nivel de abstracción para alcanzar una solución viable y efectiva, con resultados menos cuantificables y que requerirían más tiempo para lograrse. Así, la solución elegida es efectiva a corto plazo y mensurable, aunque no aborda problemas estratégicos más amplios que podrían requerir atención a largo plazo.

Además, es crucial evaluar la robustez de la solución ante cambios en los patrones de pago. Aunque la selección del modelo es efectiva actualmente, podría perder precisión con datos de usuarios con una mayor variación. Lo anterior se podría prevenir implementando un sistema de monitoreo continuo y análisis para mantener actualizado el modelo con los nuevos datos.

## 8.2. Objetivos

Analizar el cumplimiento de los objetivos específicos es crucial, ya que determina el cumplimiento del objetivo general del trabajo de título. Por ello, se estudia su completitud de forma individual, tomando en cuenta que al completar la totalidad de objetivos específicos, el objetivo general también sería completado.

### 8.2.1. Análisis de datos de usuarios

Para el análisis de los datos, se utilizó la prueba t de Student con un valor de 0,05. Aunque autores como Nigam (2022) la recomiendan para muestras con más de 30 observaciones, otros, como Banerjee (2023), prefieren la prueba Z para muestras grandes, siempre que se conozca la desviación de la población, lo cual no es el caso. Una alternativa podría ser la prueba U de Mann-Whitney, mencionada por Molina (2022).

Debido a las condiciones del trabajo, el uso de la prueba t de Student fue adecuado, aunque el valor de corte de 0,05 también podría ser discutido, ya que define cuán estricta es la diferenciación de distribuciones. Este valor, comúnmente aceptado por los investigadores, corresponde a un nivel de significancia suficientemente improbable para que no sea producto

del azar, aunque también es aceptable un valor de 0,01 (Gutiérrez, 2012).

### 8.2.2. Revisión de literatura

Definir un estándar para cumplir con este objetivo es un desafío, ya que la gran cantidad de información disponible hace imposible su revisión completa. Aunque se llevó a cabo un análisis exhaustivo de artículos relevantes para el proyecto, ciertas limitaciones técnicas y de tiempo impidieron una revisión completamente integral o la inclusión total de su contenido en el desarrollo del trabajo. No obstante, hay que notar que durante la revisión de la literatura, no se encontraron estudios nacionales que abordaran proyectos similares, lo que sugiere que este trabajo introduce una innovación en este campo. En particular, con respecto al uso de modelos de ensamble basados en múltiples variables de cada usuario.

### 8.2.3. Diseño y desarrollo de modelos

La elección de los algoritmos para la construcción de los modelos, tal como se menciona en la sección de marco conceptual, se basa en el rendimiento que estos ofrecen al clasificar. En este contexto, se podría argumentar que se excluyeron de los análisis otros algoritmos, como Support Vector Machine y Naive Bayes, entre otros mencionados por Alam (2022) en *Top 20 classification algorithms in Machine Learning*, así como las redes neuronales, debido a sus requerimientos computacionales. Sin embargo, comparar esta amplia variedad de algoritmos resulta ineficiente. Por ello, se optó por considerar los algoritmos más destacados según agencias reconocidas de gestión de riesgo, como S&P, que posiciona a Random Forest como uno de los mejores para analizar el riesgo crediticio (Vidovic y Yue, 2020). De igual manera, el Banco de Pagos Internacionales respalda el uso de algoritmos de potenciación de gradiente en este contexto (Petropoulos et al., 2019).

Otro punto que se podría modificar dentro del diseño de los modelos corresponder a la librería utilizada para optimizarlos. Este punto cobra especial importancia al considerar que los parámetros de un modelo pueden afectar en gran manera los resultados obtenidos. Para el desarrollo de esta memoria, se utilizó Optuna, una de las más conocidas y validadas por los desarrolladores; sin embargo, existen otras que también presentan resultados favorables al realizar tareas de optimización, como Hyper-Opt (Shekhar, Bansode, y Salim, 2021), aunque poseen una menor cantidad de documentación.

Para el desarrollo de este objetivo se necesitó ocupar más tiempo del esperado, debido a que la preparación de los datos resultó ser inicialmente más compleja de lo previsto. Sin embargo, con la ayuda de las librerías se pudo desarrollar tal parte de una forma más directa. Por otro lado, la fase de modelamiento también requirió un plazo mayor de desarrollo, ya que en un principio no se tenía contemplado realizarla en 3 fases, y al dividirlas, quedó el recuento sobre los 100 modelos. Un apoyo para alcanzar a completar esta tarea fue la ayuda del equipo y la paralelización de los trabajos, es decir, mientras los modelos ocupaban tiempo para ejecutarse, se realizaban otras tareas necesarias.

### 8.2.4. Preparación y validación

Para la validación de los resultados, se utilizó la técnica de validación cruzada *K-fold*, en términos simples, consiste en dividir la base de datos y alternar los grupos de entrenamiento y prueba para evaluar el rendimiento del modelo a partir de estas variaciones

(Copete, 2023). Esta metodología es ampliamente utilizada porque permite obtener un rendimiento promedio del modelo, ayudando a prevenir problemas de sobreajuste y sesgo en las predicciones. Sin embargo, un inconveniente es que este enfoque no captura adecuadamente las tendencias temporales, lo que podría ser problemático en un análisis usuario a usuario, ya que se podría perder información sobre compras consecutivas.

Por otro lado, se realizó una validación de resultados con el equipo de BNPL, en el cual se mostró el desarrollo y las conclusiones, llegando a un acuerdo de que efectivamente se mejoraban las predicciones. Se discutió también sobre lo que resultaba mejor comercialmente para los objetivos de la empresa, como la conveniencia de añadir los costos de obtención de las variables externas al modelo, lo que resultó en una respuesta afirmativa. Por ende, el modelo propuesto resultó validado cualitativamente y cuantitativamente.

## 8.3. Alcances

Los alcances del proyecto corresponden a puntos que quedaron fuera su desarrollo, ya sea por limitaciones técnicas, de recursos o temporales. En caso de no haber estado presentes estas limitaciones durante la realización de la memoria, se presupone que se podrían haber alcanzado otros resultados. También se discute su aplicabilidad en otros contextos.

- **Técnicos:** Las limitaciones técnicas se deben a la falta de conocimientos y herramientas, principalmente en programación y desarrollo web, que podrían haber sido útiles para el proyecto, pero que no se emplearon por desconocimiento del equipo y/o del estudiante. Estas limitaciones están estrechamente vinculadas con las de tiempo, ya que, por ejemplo, con más tiempo, se podría haber investigado cómo implementar la solución internamente en la web de Cleo, en lugar de externalizarla.

  Además, la formación del estudiante influyó en la capacidad para cumplir los objetivos. La familiaridad con la creación de modelos y el trabajo en Python, adquirida a través de cursos relacionados con marketing durante la carrera, permitió enfrentar de manera eficiente varios de los desafíos técnicos del proyecto.

- **De recursos:** Se refieren a la limitación de los datos disponibles en las bases de datos de Cleo. Aunque contienen los datos necesarios para que el modelo funcione correctamente, existe la interrogante sobre si estos datos también son suficientes, ya que pueden quedar variables relevantes fuera de la muestra, como cuentas bancarias secundarias o deudas no adscritas en la CMF, lo que presumiblemente podría mejorar los resultados al tener un mayor conocimiento de las personas que solicitan los créditos.

  En este sentido, aunque los datos no presentaron mucho ruido, su calidad podría haber afectado los resultados. Por ejemplo, algunos datos históricos carecían de variables presentes en el último año, lo que limitó su uso. Además, cambios en el límite de entrega de dinero de los créditos pudieron restringir el análisis.

  Otra limitación de recursos corresponde a la capacidad de procesamiento de la computadora utilizada para realizar los análisis y modelos, afectando así la disponibilidad de modelos a utilizar, excluyendo aquellos con un requerimiento computacional elevado, que se presume podrían haber entregado mejores resultados.

- **Temporales:** Corresponden a las fechas límite para la práctica y la entrega del trabajo, impidiendo desarrollar más iteraciones con otros modelos, extender el modelo a otros

servicios de Cleo, seleccionar más criterios de evaluación o elegir de manera diferente las variables a utilizar. La elección de metodología y solución también se vio afectada por el plazo, priorizando opciones que pudieran completarse en el tiempo estipulado.

- **Aplicabilidad:** La aplicación del modelo realizado presupone la obtención de datos privados, como datos bancarios de la persona, para lo cual se requiere su consentimiento y claves de acceso. Así, sería complejo extender de forma directa el modelo a otros contextos u organizaciones, a menos que también sean capaces de obtener este tipo de datos, en cuyo caso podría resultar un modelo útil para conocer el riesgo de los clientes.

## 8.4. Metodología

Si bien CRISP-DM corresponde a una de las metodologías más utilizadas en ciencia de datos y, por lo tanto, una de las más verificadas sobre su funcionamiento, según algunos autores como Salazar-Salazar et al. (2023), esta metodología, a pesar de ser completa y descriptiva, carece de roles y asignación sobre las personas que trabajan bajo su alero. Por esto, se proponen otras metodologías ágiles como Scrum-Extreme Programming o Team Data Science Process. Sin embargo, estas metodologías también presentan el defecto de necesitar un equipo suficientemente numeroso para que los roles de trabajo no sean acumulados por las mismas personas, problema que CRISP-DM no presenta.

Así, la elección de CRISP-DM fue adecuada dada la estructura del equipo y la naturaleza del proyecto, y en efecto, resultó útil para la organización de las actividades de manera general a través de los 6 pasos que la componen. Un punto débil durante el desarrollo fue la falta de directrices específicas dentro de estas fases y el desconocimiento del punto en el que detener las iteraciones.

En este sentido, un cambio que se podría haber realizado para obtener un desarrollo más directo sería desde un principio obtener directrices sobre un nivel de suficiencia de mejora para las iteraciones, en vez de ocupar la mayor cantidad posible en el tiempo determinado. Esto no solo requeriría una mayor cantidad de tiempo, sino que también generaría desgaste en el equipo. También se podrían haber definido al inicio los pasos a realizar dentro de cada fase para no extenderlos de forma innecesaria.

## 8.5. Resultados

Los datos y resultados obtenidos en este trabajo son consecuencia de los procedimientos y decisiones tomadas durante la investigación. En este sentido, se podría argumentar que cualquier cambio en el proyecto, ya sea en las acciones del estudiante, los plazos estipulados, la información disponible o incluso el conocimiento previo en ciencia de datos, podría alterar los resultados. Esto sería un símil a la teoría del caos postulada por Edward Lorentz a mediados del siglo XX, conocida popularmente como el efecto mariposa (Dizikes, 2011).

Aunque existen diferentes formas de realizar las distintas fases del proyecto, los resultados obtenidos con los pasos expuestos a lo largo del informe permitieron desarrollar un modelo que mejoró significativamente la tasa de error en comparación con el modelo actual, cumpliendo así el objetivo del trabajo. Las mejoras incluyeron una reducción significativa de la tasa de dinero entregado a falsos positivos, pasando de un 8,476 % a 0 %,

y un aumento del AUC de 0,5 a 0,962. Además, se lograron resultados positivos en términos de utilidad y rentabilidad, pasando de pérdidas de $738.409 a ganancias de hasta $1.240.673 de forma anualizada.

Con respecto al beneficio de implementación del proyecto, mostrado en la Ecuación 7.1, se prevé que el beneficio supere al costo adicional incurrido, lo que convierte esta opción en una decisión favorable desde el punto de vista empresarial. Además, este beneficio podría aumentar aún más si se considera el crecimiento constante de las compras en la población.

La implementación del modelo propuesto mejoraría la utilidad del servicio y contribuiría a su sostenibilidad a corto plazo, reduciendo significativamente el porcentaje de pérdidas y, a largo plazo, generando ganancias. Dado que BNPL es un servicio secundario en Cleo en términos de ingresos, esta mejora podría aumentar su importancia relativa en comparación con otros servicios de la compañía y su potencial para expandirse a otros negocios. No obstante, es crucial realizar evaluaciones periódicas del rendimiento del modelo, ya que valores atípicos fuera del rango de entrenamiento podrían no ser gestionados óptimamente. Esto se podría prevenir mediante el reentrenamiento regular del modelo con nuevos datos, garantizando que se mantenga actualizado y capaz de adaptarse a cambios de los usuarios.

También se realizó un análisis de sensibilidad para observar el impacto de las variables en los resultados de los modelos. Por otro lado, se investigó la utilización de variables externas de la empresa de seguridad, las cuales mejoraron el rendimiento de los modelos, sugiriendo la posibilidad de obtener otras variables externas que puedan mejorar aún más las predicciones.

## 8.6. Proyectos futuros

En relación con los resultados obtenidos y el trabajo realizado, posibles extensiones podrían ir desde el reentrenamiento del modelo con nuevos datos, es decir, no modificar la estructura del modelo sino solo los datos, ya que el comportamiento de las personas es cambiante, hasta la creación de nuevos modelos más complejos y con otros enfoques, como redes neuronales, o la utilización de los datos recabados para fomentar otros servicios. También una extensión sería añadir otro tipo de datos de fuentes externas, como variables macroeconómicas o más información de los usuarios, ya que se demostró que al menos las de la empresa de seguridad mejoraron los resultados.

En relación con los nuevos servicios, el resultado de este proyecto podría contribuir a ampliar la cartera de Cleo partiendo desde la utilización de este servicio para compras presenciales, la modificación del *output*, por ejemplo, considerando el cambio de la aceptación y denegación de pagos a la entrega de créditos de libre uso, o incluso el servicio de venta de información para los comercios, es decir, transferir el conocimiento de los datos y de los buenos y malos clientes para contribuir a la personalización de ofertas.

No obstante, estas extensiones se basan en la expectativa de un crecimiento continuo del producto en Chile. Aunque BNPL ha demostrado ser exitoso en Europa, su lugar de origen, se observa que al adaptar el servicio a las necesidades locales, tiene un gran potencial para expandirse en el continente. Un indicio de esto es la reciente alianza entre Amazon y la empresa latinoamericana Kueski de BNPL, lo que sugiere que empresas reconocidas están promoviendo activamente el desarrollo de este tipo de servicios en la región.

# 9. Conclusiones

El objetivo principal de esta memoria fue diseñar un nuevo modelo de riesgo crediticio para Cleo Chile, reduciendo la tasa de error respecto al modelo actual. Siguiendo la metodología CRISP-DM, se logró una implementación estructurada y cíclica, desde el análisis de datos hasta la mejora continua de los modelos. Como resultado, se redujo la tasa de no pago de los usuarios del 8,5 % al 0,7 % y se consiguió una utilidad positiva para el servicio. Además, se documentó un modelo mejorado con variables adicionales y se traspasaron los archivos al equipo, cumpliendo los objetivos específicos y el general.

El nuevo modelo mejora la precisión en la predicción de riesgos y proporciona información valiosa para estrategias de segmentación y desarrollo de nuevos productos. Su implementación fortalecería la posición de Cleo en el mercado, mejorando su sostenibilidad y la confianza de los clientes. Este modelo tiene el potencial de ser aplicado en otras empresas del sector financiero y en retail, ya que su enfoque adaptable y robusto permite optimizar la toma de decisiones mediante un mayor conocimiento de los clientes, reduciendo el riesgo y personalizando ofertas de crédito. Su versatilidad se debe a que utiliza técnicas avanzadas que pueden ajustarse a diferentes contextos y necesidades del mercado.

Durante el proyecto, se evaluaron diversos modelos de machine learning, incluyendo Random Forest, XGBoost y LightGBM. Los resultados mostraron que, en términos de rendimiento individual, Random Forest y LightGBM fueron los más efectivos. Sin embargo, XGBoost superó a ambos al implementarse en un esquema de modelos compuestos, donde su capacidad para ensamblar múltiples algoritmos permitió capturar patrones más complejos y mejorar la precisión global. Este enfoque de ensamble fue clave para optimizar el modelo de riesgo, logrando una predicción robusta.

Uno de los principales desafíos fue la fase de preparación de datos, que requirió limpiar y organizar minuciosamente las variables y parámetros. La falta de uniformidad en los datos iniciales exigió un esfuerzo considerable de tiempo para garantizar su calidad y consistencia. Además, la fase de modelado presentó desafíos significativos en la selección y ajuste de modelos, implicando un proceso iterativo con cambios drásticos en los resultados, dificultado por el límite temporal del proyecto para definir un modelo óptimo.

Para futuros proyectos, se recomienda incluir otras variables externas y considerar modelos más complejos, como redes neuronales, que podrían mejorar la capacidad predictiva. También se sugiere evaluar el impacto a largo plazo del modelo implementado y su aplicabilidad en diferentes mercados.

En conclusión, el trabajo colaborativo y el apoyo del equipo de BNPL fue fundamental para lograr los objetivos. Las reuniones regulares y el feedback recibido permitieron ajustar y mejorar el desempeño de los modelos de forma iterativa. Además, se aprendieron nuevas herramientas para el diseño y empaquetado de los modelos, facilitado por cursos de marketing que abordaban ciencia de datos. Así, el proyecto no solo cumplió con los objetivos planteados, sino que también proporcionó una valiosa experiencia de aprendizaje y colaboración.

# Bibliografía


Abdou, H. A., y Pointon, J. (2011, Abril). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *International journal of intelligent systems in accounting, finance & management*, *18*(2-3), 59–88. Descargado de https://doi.org/10.1002/isaf.325 doi: 10.1002/isaf.325 17

Akiba, T., Sano, S., Yanase, T., Ohta, T., y Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. En *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Francisco, CA, USA. Descargado de https://dl.acm.org/doi/10.1145/3292500.3330701 doi: 10.1145/3292500.3330701 17

Alam, B. (2022, 28 de Diciembre). Top 20 classification algorithms in Machine learning. *Medium*. Descargado de https://medium.com/@bashiralam185/top-20-classification-algorithms-in-machine-learning-e40d9dda2461 31

Alameda, T. (2019, 08 de Mayo). ¿Qué es la 'tokenización' de los pagos? *BBVA Noticias*. Descargado de https://www.bbva.com/es/que-es-la-tokenizacion-de-los-pagos/ 2

Alonso, A., y Carbó, J. M. (2021, 27 de Enero). Understanding the performance of machine learning models to predict credit Default: A Novel approach for supervisory Evaluation. *Banco de Espana Working Paper*, *2105*. Descargado de https://doi.org/10.2139/ssrn.3774075 doi: 10.2139/ssrn.3774075 14

Banerjee, C. (2023, 06 de Marzo). When to use t-test,z-test,anova and chi test? - Chandradip Banerjee - Medium. *Medium*. Descargado de https://medium.com/@chandradip93/when-to-use-t-test-z-test-anova-and-chi-test-310fd242ca62 30

Barrios Arce, J. I. (2022, 14 de Junio). *Light GBM vs XGBoost . ¿Cuál es mejor el algoritmo ?* JuanBarrios. Descargado de https://www.juanbarrios.com/light-gbm-vs-xgboost-cual-es-mejor-algoritmo/ 15

Bastos, R. (2020, 13 de Octubre). Credit Risk Analysis with Machine Learning. *Towards Data Science*. Descargado de https://towardsdatascience.com/credit-risk-analysis-with-machine-learning-736e87e95996 15

Berstein Jáuregui, S. (2024, 15 de Marzo). *Ley fintec desafíos de la implementación.* CMF. Descargado de https://www.cmfchile.cl/portal/principal/613/articles-79433_doc_pdf.pdf 2, 3

Blancas, E. (2022, 29 de Noviembre). Stop using 0.5 as the threshold for your binary classifier. *Towards Data Science*. Descargado de https://towardsdatascience.com/stop-using-0-5-as-the-threshold-for-your-binary-classifier-8d7168290d44 13

Breiman, L. (2001, Octubre). Random Forests. *Machine learning*, *45*(1), 5–32. Descargado de https://doi.org/10.1023/a:1010933404324 doi: 10.1023/a:1010933404324 14

Cardozo, P. (2023, 22 de Noviembre). *Fintech en Latinoamérica, perspectivas para el 2024 y el futuro.* Finte Chile. Descargado de https://www.fintechile.org/noticias/fintech-en-latinoamerica-perspectivas-para-el-2024-y-el-futuro 4

Chen, T., y Guestrin, C. (2016). Xgboost: A scalable tree boosting system. En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). San Francisco, CA, USA. Descargado de https://dl.acm.org/doi/10.1145/2939672.2939785 doi: 10.1145/2939672.2939785 14

Christiansen, A. (2022, 06 de Junio). *Las Fintech en Chile ya comienzan a ver a América Latina para expandirse.* La Tercera. Descargado de https://www.latercera.com/laboratoriodecontenidos/noticia/las-fintech-en-chile-ya-comienzan-a-ver-a-america-latina-para-expandirse/DWOBUMKVMRBOFMYMZDSUDWERFU/ 4

Copete, L. A. (2023, 27 de Julio). Como implementar un K-Fold Cross Validation a modelos de Machine Learning con Scikit Learn — Python. *Medium.* Descargado de https://medium.com/@luiscopete/como-implementar-un-k-fold-cross-validation-a-modelos-de-inteligencia-artificial-con-scikit-learn-eb0726c5ba55 32

Cowan, K. (2022, Mayo). *Proyecto de Ley de Innovación Financiera.* CMF. Descargado de https://www.cmfchile.cl/portal/principal/613/articles-51084_doc_pdf.pdf 3

De Losada, F. (2024, 26 de Febrero). *Transformando el futuro financiero: Tendencias fintech para 2024 y más allá.* LinkedIn. Descargado de https://www.linkedin.com/pulse/transformando-el-futuro-financiero-tendencias-fintech-de-losada-dpwke/ 4

Dizikes, P. (2011, 22 de Febrero). When the butterfly effect took flight. *MIT Technology Review.* Descargado de https://www.technologyreview.com/2011/02/22/196987/when-the-butterfly-effect-took-flight/ 33

Donges, N. (2023, 15 de Agosto). *4 Disadvantages of neural networks.* Built In. Descargado de https://builtin.com/data-science/disadvantages-neural-networks 11

Espinosa-Zúñiga, J. J. (2020, Julio). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología*, *21*(3), 1–16. Descargado de https://doi.org/10.22201/fi.25940732e.2020.21.3.022 doi: 10.22201/fi.25940732e.2020.21.3.022 14

Firican, G. (s.f.). *The history of Machine Learning.* LightsOnData. Descargado de https://www.lightsondata.com/the-history-of-machine-learning/ 12

Galli, S. (2020, 14 de Junio). Feature-engine: A new open source Python package for feature engineering. *Medium.* Descargado de https://trainindata.medium.com/feature-engine-a-new-open-source-python-package-for-feature-engineering-29a0ab88ea7c 17

Gil Martínez, C. (2020, Abril). *Interpretación de predicciones de modelos black box.* RPubs. Descargado de https://rpubs.com/Cristina_Gil/interpretacion_predicciones_modelos_bb 13

Gupta, S., y Gupta, A. (2019). Dealing with Noise Problem in Machine Learning Datasets: A Systematic Review. *Procedia Computer Science*, *161*, 466–474. Descargado de https://www.sciencedirect.com/science/article/pii/S1877050919318575 doi: 10.1016/j.procs.2019.11.146 58

Gutiérrez, P. M. (2012). p < 0,05, ¿Criterio mágico para resolver cualquier problema o leyenda urbana? *Universitas Scientiarum*, *17*(2), 203–215. Descargado de https://www.redalyc.org/pdf/499/49924592007.pdf 31

Hadri, W. (2021, 07 de Marzo). Machine learning, data science and artificial intelligence.

*Medium*. Descargado de https://medium.com/geekculture/machine-learning-data-science-and-artificial-intelligence-a45a2ffe9639 12

Hossin, M., y Sulaiman, M. (2015, Marzo). A review on Evaluation Metrics for Data Classification Evaluations. *International journal of data mining and knowledge management process*, *5*(2), 01–11. Descargado de https://doi.org/10.5121/ijdkp.2015.5201 doi: 10.5121/ijdkp.2015.5201 15

Howarth, J. (2024, 25 de Enero). *27 Buy Now, Pay Later Statistics (2024 & 2025)*. Exploding Topics. Descargado de https://explodingtopics.com/blog/bnpl-stats 43

Kalimi, A. (2023, 16 de Agosto). List of machine learning models. *Medium*. Descargado de https://medium.com/@codekalimi/list-of-machine-learning-models-61b51ad492f1 13

Kotsiantis, S., Kanellopoulos, D., y Pintelas, P. (2007, 28 de Diciembre). Data preprocessing for supervised leaning. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *1*(12), 4104–4109. Descargado de https://publications.waset.org/14136/pdf doi: 10.5281/zenodo.1082415 49

Kroese, D., Botev, Z., Taimre, T., y Vaisman, R. (2019). *Data science and machine learning: Mathematical and statistical methods*. Boca Raton: CRC Press. Descargado de https://people.smp.uq.edu.au/DirkKroese/DSML/ 12

Kryńska, K. (2020, 05 de Julio). *Comparison of models for credit risk purposes - logistic regression vs random forest.* RPubs. Descargado de https://rpubs.com/kkrynska/CreditScoring 14

Kumar, S. (2024, 11 de Enero). *Balancing Act: The Pros and Cons of Machine Learning Algorithms.* LinkedIn. Descargado de https://www.linkedin.com/pulse/balancing-act-pros-cons-machine-learning-algorithms-mba-ms-phd-aty6c/ 14

Lizares, M. (2017). *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico* (Tesina de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística, Lima, Perú). Descargado de https://cybertesis.unmsm.edu.pe/handle/20.500.12672/7122 14

Lorenzo, J. (2019). Estadística básica: Introducción a la prueba t de student y el análisis de la varianza. *Ansenuza*. Descargado de https://ansenuza.ffyh.unc.edu.ar/bitstream/handle/11086.1/1348/Prueba%20t%20y%20ANOVA.pdf 15

Maldonado Rosas, D. (2021, 10 de Agosto). *Qué es el Open Banking y por qué compartir nuestra información revolucionaría al sistema financiero.* Fintualist. Descargado de https://fintualist.com/chile/economia/que-es-el-open-banking-y-por-que-compartir-nuestra-informacion-revolucionaria-al-sistema-financiero/ 22

Mankad, S. (2020, 01 de Diciembre). *A tour of evaluation Metrics for Machine Learning.* Analytics Vidhya. Descargado de https://www.analyticsvidhya.com/blog/2020/11/a-tour-of-evaluation-metrics-for-machine-learning 10

Marmuzevich, S. (2023, 07 de Diciembre). *¿Qué es el iGaming?* Sowtswiss. Descargado de https://www.softswiss.com/es/knowledge-base/igaming-definition/ 5

Martinez, C. V. (2024, 08 de Marzo). Economista Víctor Salas y alza del IPC: "Es una señal

de que no tenemos controlada la inflación". *Diario Uchile*. Descargado de https://radio.uchile.cl/2024/03/08/economista-victor-salas-y-alza-del-ipc-es-una-senal-de-que-no-tenemos-controlada-la-inflacion/ 8

Mathur, G. (2023, 06 de Julio). *Data science vs. machine learning: What's the difference?* IBM. Descargado de https://www.ibm.com/blog/data-science-vs-machine-learning-whats-the-difference/ 12

Mena Álamos, J. (2023, 17 de Agosto). *El boom tecnológico llegó para quedarse: La evolución de los métodos de pago.* LinkedIn. Descargado de https://www.linkedin.com/pulse/el-boom-tecnolÃşgico-llegÃş-para-quedarse-la-evoluciÃşn-de-los-mÃľtodos/ 2

Mercado, P. (2022, 22 de Abril). El papel clave de las 'fintech' en el camino hacia la sostenibilidad. *Cinco Días*. Descargado de https://cincodias.elpais.com/cincodias/2022/04/22/opinion/1650637861_170064.html 2

Mina Chiquiza, A. R. (2024, 19 de Enero). *Técnicas avanzadas de aprendizaje automático: ensamblaje de modelos, aprendizaje profundo, procesamiento de lenguaje natural.* LinkedIn. Descargado de https://www.linkedin.com/pulse/tÃľcnicas-avanzadas-de-aprendizaje-automÃątico-modelos-mina-chiquiza-pnmue/ 14

Minoli, M. (2020, 17 de Octubre). *Adaptación de Scrum para el desarrollo de soluciones de Business Analytics con CRISP-DM.* Hiberus. Descargado de https://www.hiberus.com/crecemos-contigo/adaptacion-de-scrum-para-business-analytics/ 18

Molina, M. (2022, 13 de Junio). *Prueba de la U de Mann-Whitney. Ciencias o letras. - AnestesiaR.* Anestesiar. Descargado de https://anestesiar.org/2022/prueba-de-la-u-de-mann-whitney-ciencias-o-letras/ 30

Mondal, R. (2024, 10 de Julio). Feature transformation- part of feature engineering. *Medium*. Descargado de https://medium.com/@datasciencejourney100_83560/feature-transformation-part-of-feature-engineering-dff2deaf59a2 23

Moreno, R. C. . (2024, 04 de Enero). *Tendencias Fintech 2024.* LinkedIn. Descargado de https://www.linkedin.com/pulse/tendencias-fintech-2024-rinconcardenas-com-imtbe/ 4

Narkhede, S. (2018, 26 de Junio). Understanding AUC - ROC Curve. *Towards Data Science*. Descargado de https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 10

Nigam, V. (2022, 08 de Febrero). *Statistical tests: when to use T-Test, Chi-Square and more.* Built In. Descargado de https://builtin.com/data-science/t-test-vs-chi-square 30

Olmos, R. (2024, 14 de Marzo). Los nudos que entrampan el proyecto de ley de protección de datos personales. *Diario Financiero*. Descargado de https://www.df.cl/df-lab/transformacion-digital/los-nudos-que-entrampan-el-proyecto-de-ley-de-proteccion-de-datos 8

Peralta, F. C. (2014, 27 de Octubre). Proceso de Conceptualización del Entendimiento del Negocio para Proyectos de Explotación de Información. *Revista latinoamericana de ingeniería de software*, *2*(5), 273. Descargado de https://doi.org/10.18294/relais.2014.273-306 doi: 10.18294/relais.2014.273-306 19

Pescio, B. (2022, 02 de Noviembre). Buy Now Pay Later: la nueva tendencia entre los millennials y generación Z que se expande en Chile. *Diario Financiero*. Descargado

de https://dfmas.df.cl/df-mas/como-cuido-mis-lucas/buy-now-pay-later-la-nueva-tendencia-entre-los-millenials-y-generacion 4

Petropoulos, A., Siakoulis, V., Stavroulakis, E., y Klamargias, A. (2019). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *IFC Bulletins chapters*, *49*. Descargado de https://www.bis.org/ifc/publ/ifcb49_49.pdf 31

Ponasso, L. (2024, 09 de Marzo). A contramano: por qué cada vez más empresas del mundo online se vuelcan al físico. *La Nación*. Descargado de https://www.lanacion.com.ar/economia/negocios/a-contramano-por-que-cada-vez-mas-empresas-del-mundo-online-se-vuelcan-al-fisico-nid09032024/ 1, 2

Promueve la competencia e inclusión financiera a través de la innovación y tecnología en la prestación de servicios financieros, Ley Fintec. Ley Nº21.521. (2024). *Diario Oficial de la República de Chile*. 2

Pérez Rojas, A. R. (2016). *Diseño de metodologÍa para el seguimiento de modelos de riesgo crediticio* (Memoria para optar al título de ingeniero civil industrial, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ingeniería Industrial, Santiago, Chile). Descargado de https://repositorio.uchile.cl/handle/2250/144507 13

Rainio, O., Teuho, J., y Klén, R. (2024, 13 de Marzo). Evaluation metrics and statistical tests for machine learning. *Scientific reports*, *14*(6086). Descargado de https://www.nature.com/articles/s41598-024-56706-x doi: 10.1038/s41598-024-56706-x 15

Rosidi, N. (2023, 22 de Marzo). *Machine Learning: What is Bootstrapping?* KDnuggets. Descargado de https://www.kdnuggets.com/2023/03/bootstrapping.html 14

Salazar-Salazar, G., Mora, M., Duran-Limon, H. A., y Álvarez Rodríguez, F. J. (2023). *A selective comparative review of CRISP-DM and TDSP development methodologies for big data analytics systems*. Descargado de https://doi.org/10.1007/978-3-031-40956-1_6 doi: 10.1007/978-3-031-40956-1_6 33

Saltz, J. (2022, 02 de Mayo). *CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects*. Data Science Process Alliance. Descargado de https://www.datascience-pm.com/crisp-dm-still-most-popular/ 17

Schwartz, B. (2024, 25 de Abril). *Kanban History: origin & expansion across industries.* Proyect Manager. Descargado de https://www.projectmanager.com/blog/kanban-history 18

Shekhar, S., Bansode, A., y Salim, A. (2021, 12). A Comparative study of Hyper-Parameter Optimization Tools. *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. Descargado de https://doi.org/10.1109/csde53843.2021.9718485 doi: 10.1109/csde53843.2021.9718485 31

Talaviya, A. (2023, 30 de Octubre). CRISP-DM framework: A foundational data mining process model. *Medium*. Descargado de https://medium.com/@avikumart_/crisp-dm-framework-a-foundational-data-mining-process-model-86fe642da18c 17

Talreja, A. (2023, 19 de Junio). *Scrum Origins and Agile Principles: The Foundations of Agile Project Management.* Teaching Agile. Descargado de https://teachingagile.com/scrum/psm-1/scrum-theory-principles/scrum-history 18

Tavor, D., Wolfson, A., y Mark, S. (2013, Marzo). Are Internet-Based services more sustainable? *International journal of computer and technology*, *4*(1), 129–134. Descargado de https://doi.org/10.24297/ijct.v4i1c.3118 doi: 10.24297/ijct.v4i1c.3118 2

Tejeda, F. B. (2024, 19 de Febrero). BNPL (Compra Ahora y Paga Después): Qué es, como funciona y ventajas. *Ecommerce News*. Descargado de https://www.ecommercenews.pe/comercio-electronico/2024/bnpl-que-es.html/ 4

Thomas, S. (2022, 17 de Febrero). *Discrete vs. Continuous Variables: Differences Explained.* Outlier. Descargado de https://articles.outlier.org/discrete-vs-continuous-variables 13

Trevisan, V. (2022, 17 de Enero). Using SHAP values to explain how your machine learning model works. *Towards Data Science*. Descargado de https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137 13, 17

Urrego, N. (2023, 11 de Julio). Transformando datos en oro con R y Python: Cómo la estandarización y normalización mejoran tus resultados. *Medium*. Descargado de https://nicolasurrego.medium.com/transformando-datos-en-oro-c%C3%B3mo-la-estandarizaci%C3%B3n-y-normalizaci%C3%B3n-mejoran-tus-resultados-fbe0840d2b94 24

Velayudhan, A. (2020, 19 de Enero). *Type I and Type II errors in Credit Scoring - Need for a clear definition.* LinkedIn. Descargado de https://www.linkedin.com/pulse/type-i-ii-errors-credit-scoring-need-clear-velayudhan-frm-dipifr/ 16

Vidovic, L., y Yue, L. (2020). Machine learning and credit risk modelling. *SP Global*. Descargado de https://www.spglobal.com/marketintelligence/en/documents/machine_learning_and_credit_risk_modelling_november_2020.pdf 31

Visus, A. (2020, Octubre). *¿Para qué sirve Python? Razones para utilizar este lenguaje de programación.* ESIC University. Descargado de https://www.esic.edu/rethink/tecnologia/para-que-sirve-python 17

Vujović, Ž. (2021, Julio). Classification model evaluation metrics. *International journal of advanced computer science and applications/International journal of advanced computer science & applications*, *12*(6). Descargado de https://doi.org/10.14569/ijacsa.2021.0120670 doi: 10.14569/ijacsa.2021.0120670 15

Wakefield, K. (s.f.). *A guide to the types of machine learning algorithms and their applications.* SAS. Descargado de https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html 12

Xu, P., Ji, X., Li, M., y Lü, W. (2023, Marzo). Small data machine learning in materials science. *npj computational materials*, *9*(1). Descargado de https://doi.org/10.1038/s41524-023-01000-z doi: 10.1038/s41524-023-01000-z 15

Yurdakul, B. (2018). *Statistical properties of population stability index* (Tesis Doctoral, Western Michigan University). Descargado de https://scholarworks.wmich.edu/dissertations/3208 49

# Anexos

## Anexo A. Antecedentes

### A.1. Evolución y crecimiento de la industria

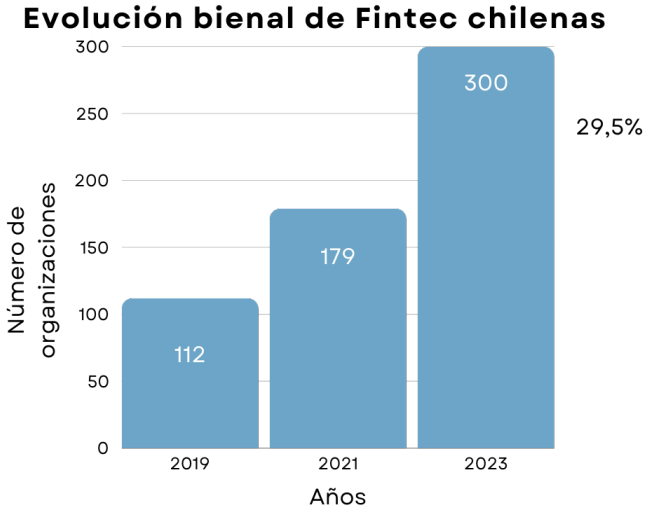


Figura A.1: Evolución bienal, o cada dos años, de Fintec chilenas. Fuente: Adaptado de *Ley Fintec Desafíos de la implementación*.- CMF.

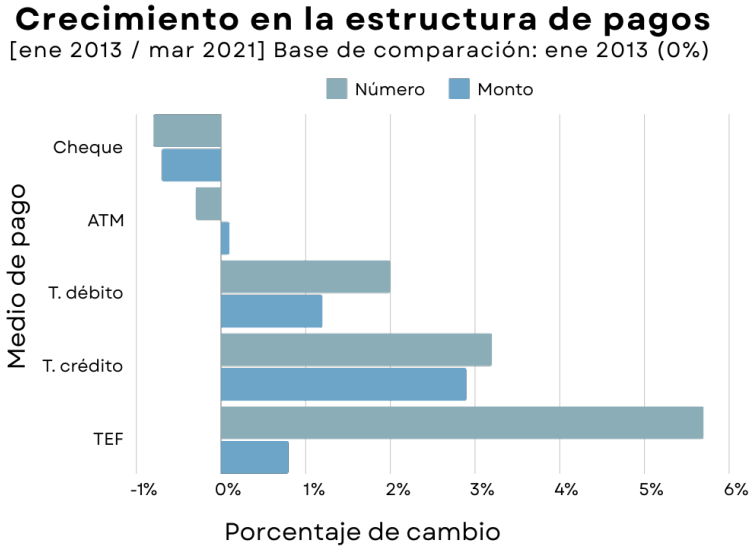


Figura A.2: Crecimiento en la estructura de pagos con referencia en valores del año 2013. Fuente: Adaptado de *Proyecto de Ley de Innovación Financiera*.- CMF.

## A.2. Comparación comisiones

| Empresa | Valor comisión comercio |
|---|---|
| Cleo | 3 % |
| Wibond | 4,79 % considerando costo promedio transaccional y financiero[a] |
| Ventipay | 3,49 %[b] |

Tabla A.1: Comparación comisiones al comercio de Cleo, Wibond y Ventipay. Fuente: *Elaboración propia*

[a] Para más información: https://ayuda.tiendanube.com/es_ES/wibond/preguntas-frecuentes-sobre-wibond
[b] Para más información: https://ventipay.com/comercios/suscripciones/#tarifas

## A.3. FODA detallado

A continuación, se presenta el análisis FODA detallado referente a los ítems de la Figura 1.2.

- **Fortalezas:** El talento y los recursos humanos (RR.HH.) de calidad son pilares fundamentales en el funcionamiento de Cleo. Como se muestra en el Anexo A.3, una gran cantidad de los colaboradores y practicantes provienen de universidades de renombre, lo que da cuenta de las capacidades y habilidades del equipo. La eficiencia en sus operaciones, respaldada por una estructura empresarial sólida, junto con su alta flexibilidad para brindar servicios personalizados a sus clientes, constituyen ventajas significativas para la empresa.

  Además, la exposición al mercado europeo, de donde proviene la compañía y donde se concentra la mayor cantidad de servicios de BNPL (Howarth, 2024), promueve la innovación y la rápida adaptación a nuevas tendencias, permitiendo la asimilación de mejores prácticas y estándares de la industria.

- **Oportunidades:** Como se mencionó en la Sección 1.2.1, la adopción de nuevas tecnologías como el pago móvil y la diversificación hacia más canales de pago son oportunidades que Cleo podría aprovechar. Asimismo, la diversificación de servicios podría reducir el riesgo y ampliar las oportunidades de crecimiento de la organización.

- **Debilidades:** Uno de los mayores desafíos que enfrenta Cleo es la limitación de recursos. Su reciente establecimiento y su tamaño reducido restringen el acceso a una mayor cantidad de recursos humanos y a inversiones significativas, lo que dificulta su competitividad frente a otros actores más establecidos en la industria *Fintech*. Otra debilidad que se presenta correspondería a las altas tasas de no pago que presenta el servicio, lo que impide mantener valores competitivos con la competencia, y por sobre todo, evita que el servicio pueda ser rentable.

- **Amenazas:** Las nuevas tecnologías y el mercado aún poco explorado en Chile representan una amenaza para Cleo. El atractivo de este mercado emergente por sus bajas barreras de entrada puede generar competencia por parte de otras organizaciones similares, lo que podría poner en peligro la posición de Cleo en el mercado y su capacidad para alcanzar sus objetivos de crecimiento. Además, el hecho de que la ley *Fintech* aún esté en proceso de cambio es un punto de riesgo para las operaciones.

## A.4. Datos sobre la organización

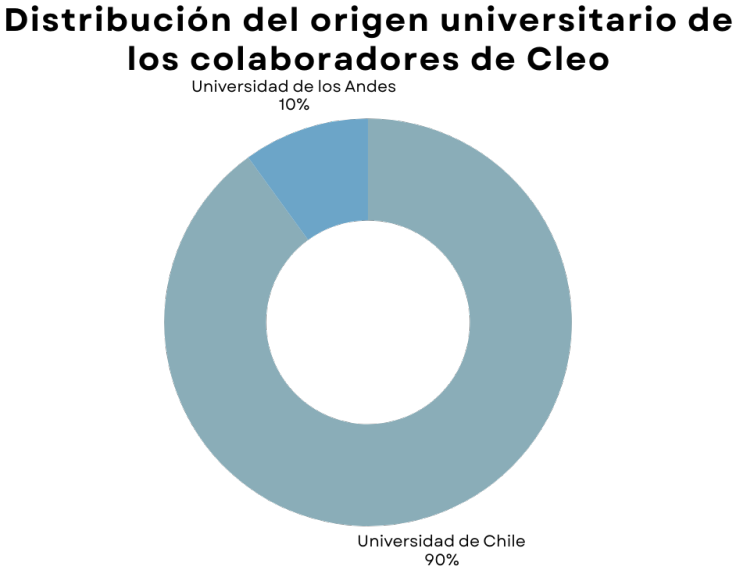


Figura A.3: Distribución del origen universitario de los trabajadores. Fuente: *Elaboración propia*

# Anexo B. Descripción del problema

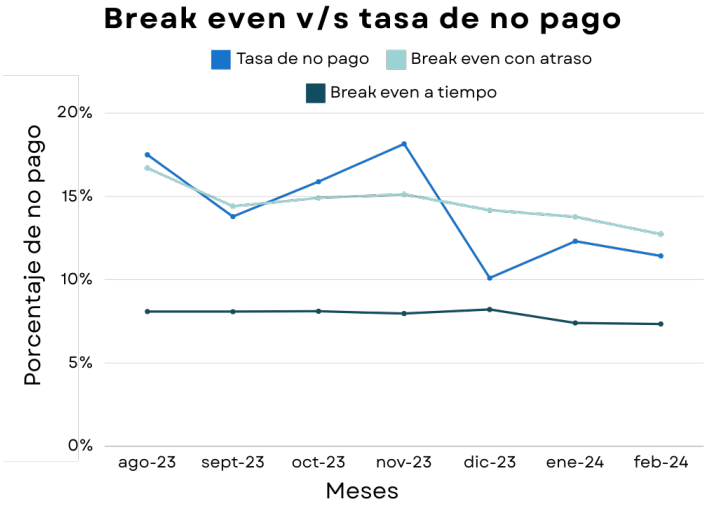


Figura B.1: Variación de *break even* a tiempo y con atraso con respecto a la tasa de no pago. Fuente: *Elaboración propia*

Con respecto al gráfico superior, el *break even* a tiempo se refiere al límite superior de la tasa de no pago para obtener utilidades con usuarios que pagan a tiempo. El *break even* con atraso es similar, solo que se agregan los usuarios que pagan con retraso.

# Anexo C. Marco teórico

## C.1. Modelos

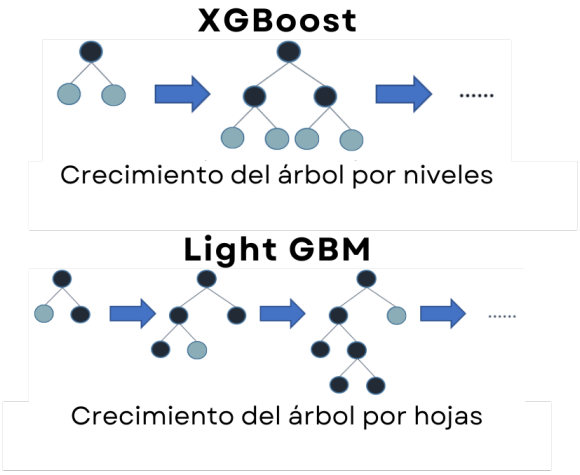


Figura C.1: Diferencia entre XGBoost y Light GBM. Fuente: Adaptado de *Light GBM vs XGBoost . ¿Cuál es mejor el algoritmo ?*.- Barrios, J.

# Anexo D. Desarrollo y resultados

## D.1. Gráficas del modelo actual

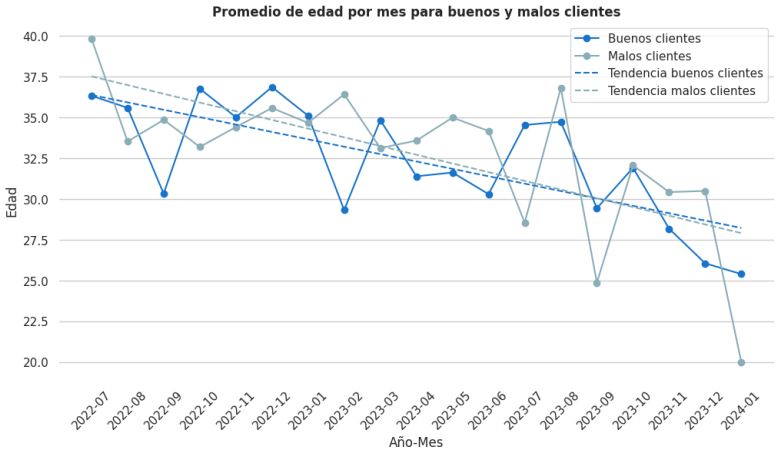


Figura D.1: Gráfica de promedio mensual de edad para buenos y malos usuarios. Fuente: *Elaboración propia*

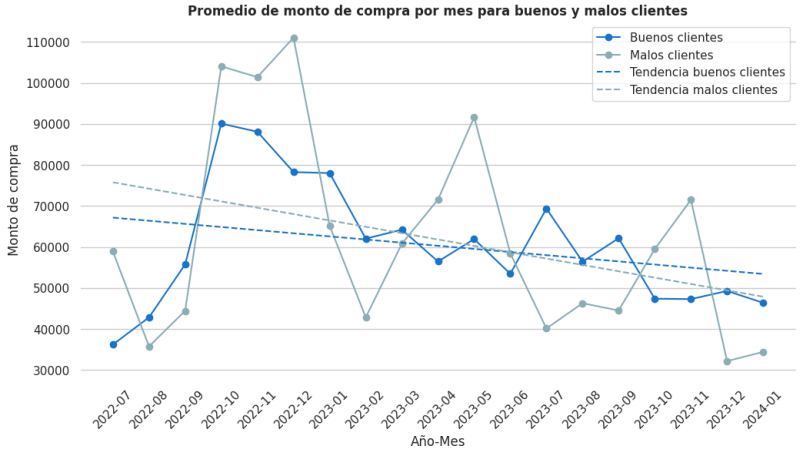


Figura D.2: Gráfica de promedio mensual de monto de compra para buenos y malos usuarios. Fuente: *Elaboración propia*

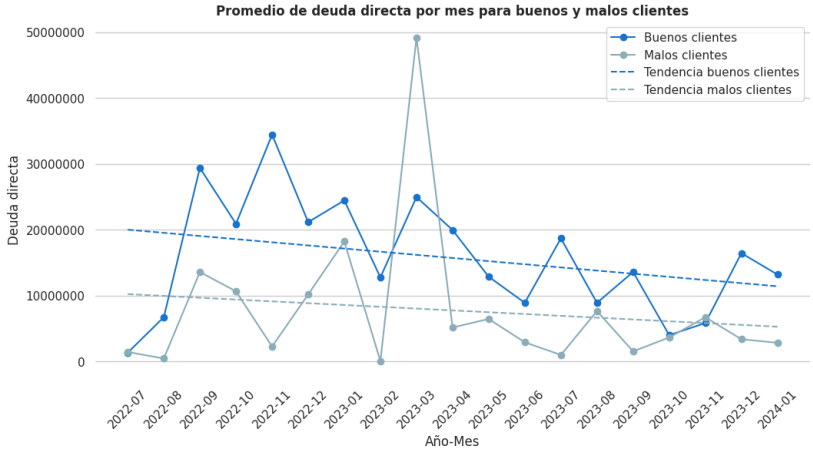


Figura D.3: Gráfica de promedio mensual de deuda directa para buenos y malos usuarios. Fuente: *Elaboración propia*

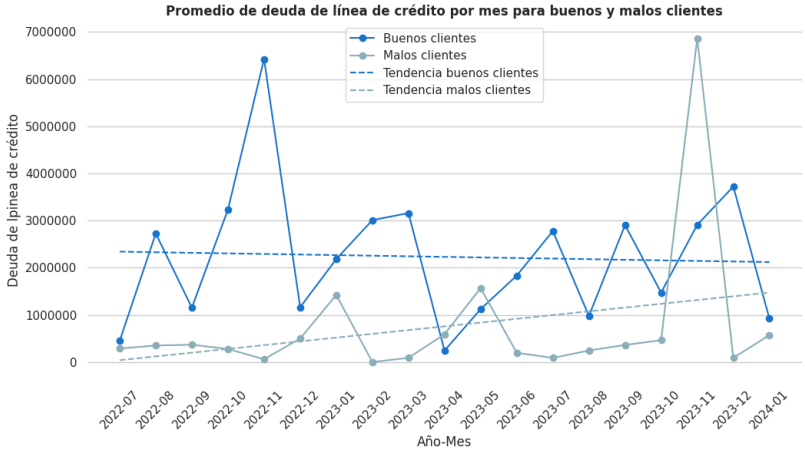


Figura D.4: Gráfica de promedio mensual de deuda de línea de crédito para buenos y malos usuarios. Fuente: *Elaboración propia*

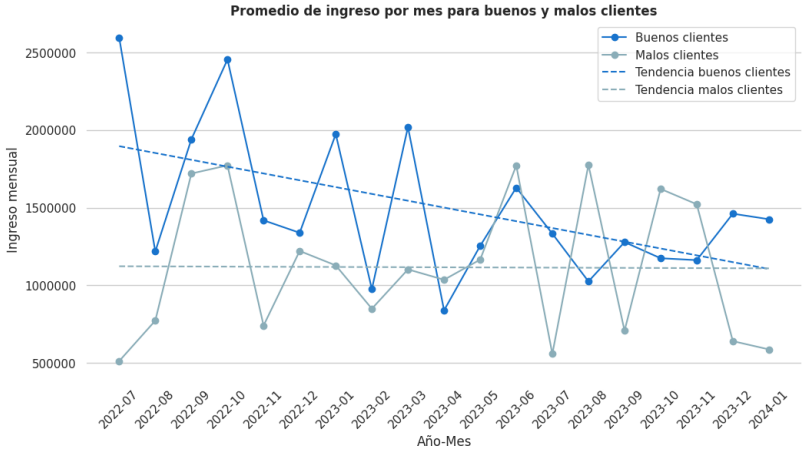


Figura D.5: Gráfica de ingreso promedio mensual para buenos y malos usuarios. Fuente: *Elaboración propia*

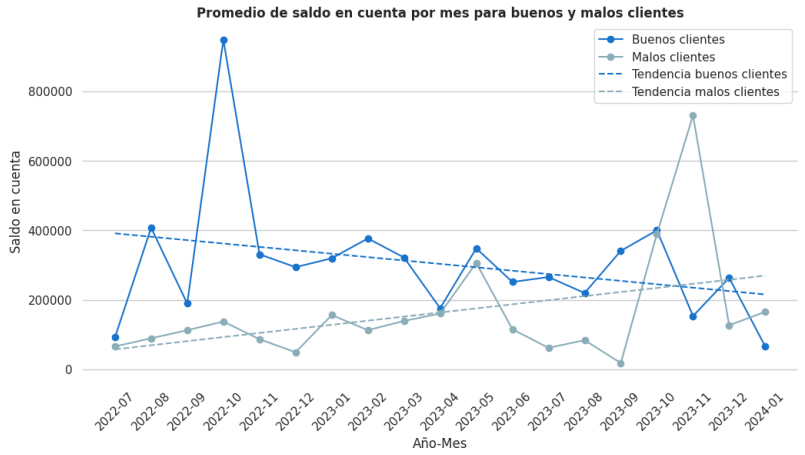


Figura D.6: Gráfica de saldo en cuenta promedio mensual para buenos y malos usuarios. Fuente: *Elaboración propia*

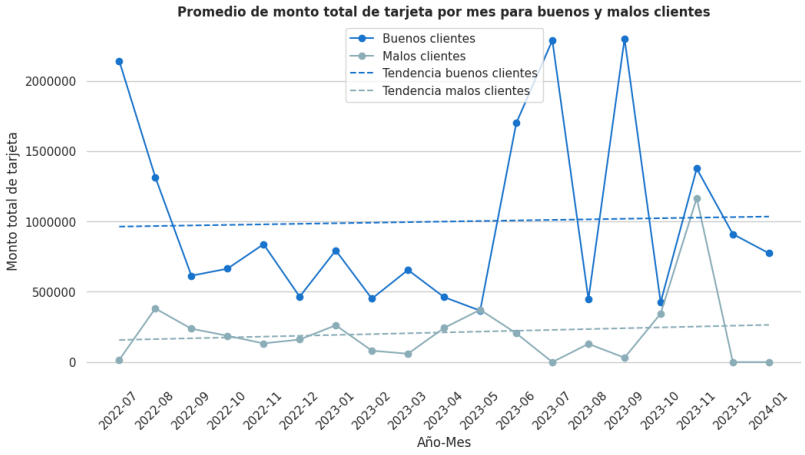


Figura D.7: Gráfica de promedio mensual de monto total de tarjeta para buenos y malos usuarios. Fuente: *Elaboración propia*

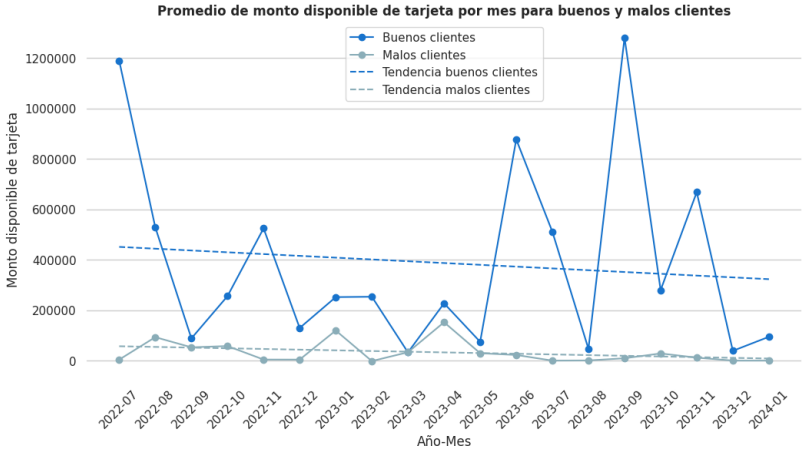


Figura D.8: Gráfica de promedio mensual de monto disponible de tarjeta para buenos y malos usuarios. Fuente: *Elaboración propia*

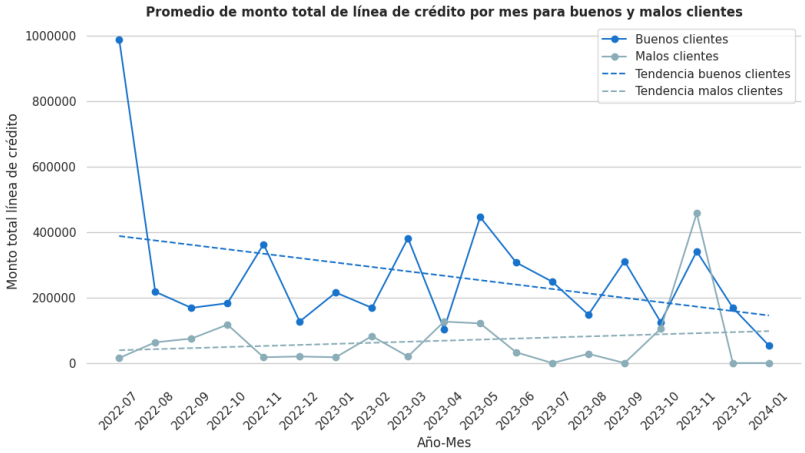


Figura D.9: Gráfica promedio mensual de monto total de línea de crédito para buenos y malos usuarios. Fuente: *Elaboración propia*

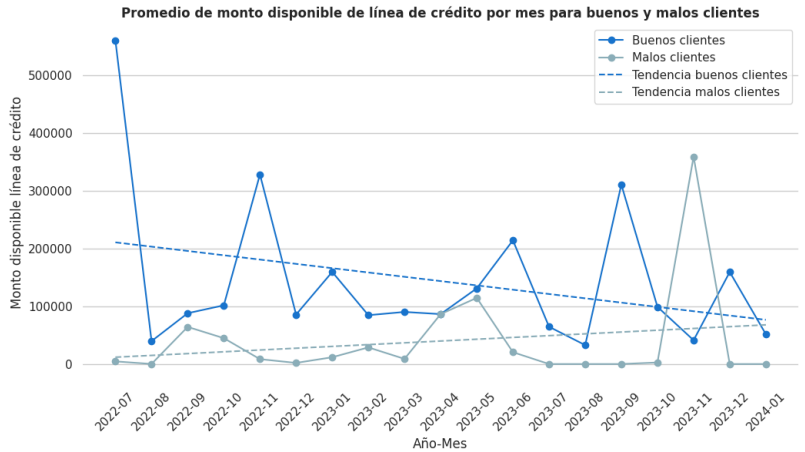


Figura D.10: Gráfica promedio mensual de monto disponible de línea de crédito para buenos y malos usuarios. Fuente: *Elaboración propia*

## D.2. Preparación de los datos detallada

Se presentan los cinco pasos descritos en la Sección 7.3 de manera más detallada, además de incluir una explicación sobre las herramientas utilizadas.

1. El primer filtro se aplica a las compras confirmadas, ya que las compras canceladas o en otro estado no son relevantes para los fines de este trabajo. Además, se excluyen las compras cuyo período de pago aún está vigente, dado que no es posible determinar si el usuario pagará o no en estos casos.

2. Las herramientas utilizadas en este paso incluyen los métodos SmartCorrelatedSelection y DropHighPSIFeatures de la librería Feature-engine. Estos métodos permiten seleccionar las variables basándose en criterios de correlación y en el índice PSI, el cual busca preservar la estabilidad del modelo seleccionando variables que no impacten significativamente al variar (Yurdakul, 2018).

3. Aunque los criterios automáticos del paso anterior ayudan a identificar las mejores variables, no son absolutos. Por ello, en este paso se revisan manualmente las variables seleccionadas para decidir cuáles conservar y cuáles eliminar. Por ejemplo, el método DropHighPSIFeatures podría seleccionar una variable relacionada con el número de celular de una persona, lo cual no tiene sentido al predecir si una persona pagará o no.

   Así, entre estos dos métodos se realiza un cruce de datos y de forma heurística se seleccionan las variables adecuadas para la predicción de los usuarios, teniendo en cuenta que el análisis de la Sección 7.2 refleja también variables significativas para este análisis. Después de escoger las variables a utilizar, se depura la base eliminando las filas que no tengan datos en las columnas referentes a estas variables seleccionadas.

4. Con las variables principales seleccionadas se crean nuevas utilizando los siguientes criterios:

   - **División:** Se dividen las variables seleccionadas referentes a dinero, como el ingreso o la deuda, en el monto de compra y las cuotas a pagar, esto con el objetivo de que estas nuevas variables reflejen el impacto de la compra en las finanzas de la persona, es decir, indicadores como capacidad de pago o cuanta deuda representaría de la que la persona tiene actualmente.
   - **Multiplicación:** Usando el método RelativeFeatures de Feature-engine, se multiplican las variables seleccionadas por variables relacionadas con la compra, como monto y cuotas, además de elevarlas al cuadrado, con el objetivo de mejorar la predicción capturando la interacción entre variables.
   - **Discretización:** A través de un método llamado EqualFrequencyDiscretiser de Feature-engine, las variables continuas son asignadas a rangos con la misma frecuencia de datos. Esto se realiza a partir de las mismas variables ya mencionadas, y según Kotsiantis, Kanellopoulos, y Pintelas (2007) este método contribuye a mejorar la predicción del modelo.

5. Luego de terminar de eliminar, crear y escoger variables, se transforman utilizando el método QuantileTransformer de Sklearn. Esta transformación se utiliza para asignar una distribución normal a las variables, mejorando así el desempeño del modelo.

## D.3. Modelado detallado

Se expande la explicación sobre las iteraciones que componen los pasos del modelado que se aprecia en la Figura 7.3.

- En la primera iteración, se utilizaron los algoritmos Random Forest, XGBoost y LightGBM con la base de datos completa. Para cada uno de estos algoritmos, se diseñaron tres modelos basados en la procedencia de los datos: uno referente a los datos bancarios y de ingresos de los usuarios, otro basado en los datos de deuda registrados en la CMF, y un tercer modelo que combina estos dos mediante el método de ensamblaje.

  La separación de las bases de datos para la fase de modelado se basa en la reducción de dimensionalidad en cada base por separado, es decir, que el conjunto de variables de cada base pueda capturar de forma adecuada las interacciones de este mismo.

- En la segunda iteración, se seleccionó el mejor algoritmo de la fase anterior y, basándose en este, se diseñaron nuevos modelos utilizando la segmentación de la base de datos. Además, se mantuvo la separación previa de la procedencia de los datos para comparar los resultados de esta fase con los del modelo actual de Cleo. Las segmentaciones utilizadas incluyeron usuarios con montos de compra por debajo y por encima del percentil 50, por debajo y por encima del percentil 70, usuarios que compran en 1, 3 y 6 cuotas, y una segmentación específica para uno de los clientes más relevantes de Cleo, Sky Airlines.

  Este tipo de segmentaciones se basa en la utilización de variables referentes a las órdenes, más que en las personas, con el objetivo de evitar cualquier sesgo de predicción. Las divisiones de montos de compra se realizan de forma equitativa (50/50) y de forma desplazada (70/30) para capturar información sobre si el comportamiento de los compradores es lineal o si se acentúan diferencias en las compras de mayor valor.

  Por su parte, las cuotas forman una parte fundamental del proceso de compra en Cleo, ya que dependiendo de ellas se genera una mayor ganancia por comisiones. Finalmente, Sky Airlines presenta el mayor número de órdenes realizadas, representando un 33 % de la totalidad de órdenes de un total de 150 compañías, por lo que cobra especial relevancia la clasificación correcta de los usuarios provenientes de ese cliente.

- Para la tercera iteración, se utilizaron las mismas segmentaciones que en la segunda iteración, pero se agregaron nuevas variables externas de una empresa de seguridad. Estas variables corresponden a puntajes basados en el riesgo de que una persona pueda cometer fraude, obtenidos a partir de patrones de navegación y comportamiento en internet. Aunque estas variables no se refieren directamente al no pago por parte de una persona, se evaluó su utilidad en esta fase del modelado. En este sentido, el modelado de esta iteración pasó por:

  – La inclusión o no de las variables de la empresa externa en los modelos de ingresos y deuda.

  – Las combinaciones de ensamblaje entre el modelo de ingresos, el modelo de deuda y un nuevo modelo basado solamente en las nuevas variables.

  – Las segmentaciones mencionadas de compras por debajo y por encima de ciertos percentiles, cuotas y comercio.

## D.4. Iteración 1

| | Bancario | | | CMF | | | Bancario + CMF | | |
|---|---|---|---|---|---|---|---|---|---|
| **Algoritmo** | **Utilidad** | **No pago** | **AUC** | **Utilidad** | **No pago** | **AUC** | **Utilidad** | **No pago** | **AUC** |
| Random Forest | $7.249.714 | 0,596 % | 0,947 | $5.192.309 | 3,079 % | 0,870 | $7.750.523 | 0,696 % | 0,954 |
| XGBoost | $5.258.174 | 0,089 % | 0,877 | $2.659.058 | 0,000 % | 0,747 | $7.574.309 | 0,141 % | 0,969 |
| LightGBM | $7.244.696 | 0,784 % | 0,941 | $5.413.569 | 2,692 % | 0,862 | $7.801.970 | 0,677 % | 0,957 |

Tabla D.1: Comparación resultados Random Forest, XGBoost y LightGBM. Fuente: *Elaboración propia*

## D.5. Iteración 2

| **Modelo** | **Segmentación** | **Utilidad** | **No pago** | **AUC** |
|---|---|---|---|---|
| Bancario + CMF | 50 / 50 | $ 723.726 | 0,741 % | 0,962 |
| Bancario + CMF | Sky / No Sky | $ 686.842 | 1,058 % | 0,941 |
| Bancario | 50 / 50 | $ 666.622 | 0,000 % | 0,929 |
| Bancario + CMF | 1 / 3 / 6 cuotas | $ 602.960 | 1,675 % | 0,921 |
| CMF | 50 / 50 | $ 596.182 | 0,000 % | 0,868 |
| Bancario | Sky / No Sky | $ 583.808 | 0,655 % | 0,902 |
| Bancario | 70 / 30 | $ 579.266 | 0,408 % | 0,894 |
| Bancario | 1 / 3 / 6 cuotas | $ 531.608 | 0,000 % | 0,853 |
| CMF | 1 / 3 / 6 cuotas | $ 494.158 | 0,272 % | 0,785 |
| CMF | 70 / 30 | $ 364.793 | 0,000 % | 0,785 |
| Bancario + CMF | 70 / 30 | $ 253.194 | 4,268 % | 0,804 |
| CMF | Sky / No Sky | $ 118.555 | 0,000 % | 0,667 |

Tabla D.2: Comparación resultados segmentaciones y modelos ordenados por utilidad de mayor a menor. Fuente: *Elaboración propia*

## D.6. Iteración 3

A partir de los resultados utilizados para calcular la Tabla 7.5, que compara el promedio de los 10 mejores modelos con y sin las variables de la empresa de seguridad, se generaron las gráficas de caja presentadas en la Figura D.11. Estas gráficas muestran las métricas de utilidad, tasa de no pago y valor AUC.

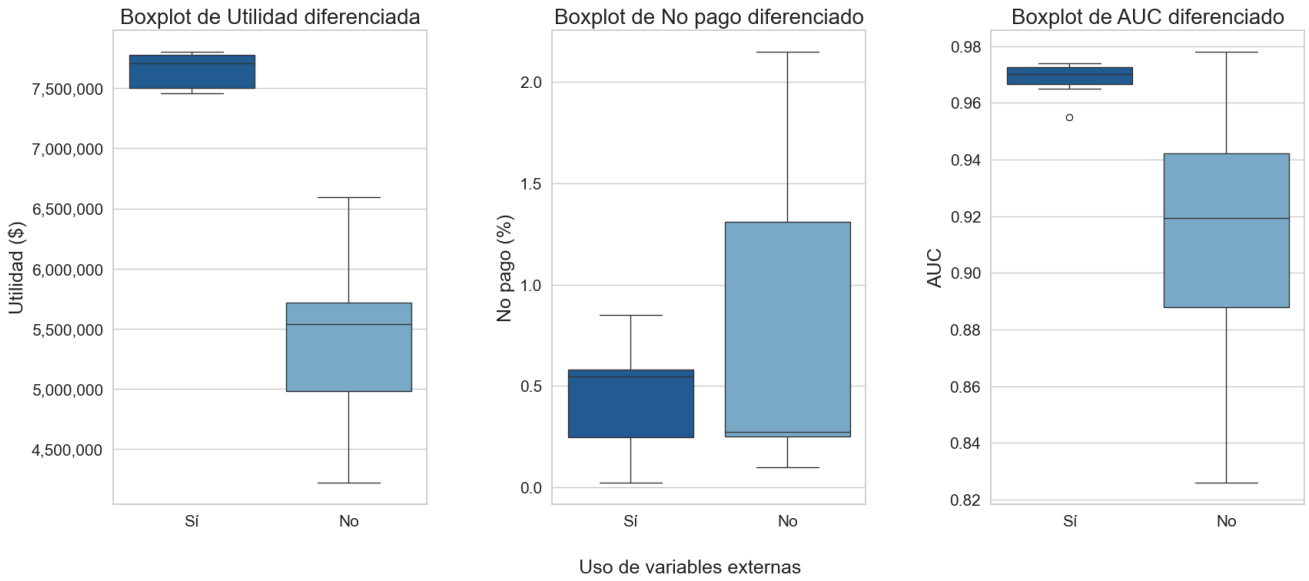


Figura D.11: Gráficos de caja en comparación para los 10 mejores resultados de modelos con y sin uso de variables externas, en dónde de izquierda a derecha corresponderían a: 1. Utilidad, 2. Tasa de no pago y 3. Valor AUC. Fuente: *Elaboración propia*

En estas gráficas, además de observar una mejora en el valor promedio, se destaca que los modelos que incorporan las variables externas muestran una mayor estabilidad, evidenciada por una menor varianza. Este hallazgo es significativo, ya que sugiere que los datos adicionales no solo mejoran la precisión de las predicciones, sino que también estabilizan los resultados, lo cual es crucial para la consistencia del modelo. No obstante, esta mayor estabilidad podría dificultar la detección de valores atípicos, que a menudo requieren modelos más flexibles.

La mejora observada también sugiere que existen comportamientos de los usuarios que no están siendo capturados por las variables y datos actuales de Cleo. Esto revela una oportunidad para explorar nuevas fuentes de información que complementen las bases de datos existentes. Algunas de estas variables adicionales podrían ser específicas del individuo, como información demográfica o patrones de comportamiento particulares. Otras podrían ser de índole macroeconómica, como cambios en el consumo de la población o el crecimiento económico del país. Además, las variables transaccionales, aún no implementadas debido a la falta de compradores recurrentes, podrían proporcionar una visión más completa y detallada del comportamiento del usuario.

| Uso variables externas | Modelo | Segmentación | Utilidad | No pago | AUC |
|---|---|---|---|---|---|
| Bancario y CMF | Bancario + CMF + Externo | Sky / No Sky | $7.802.419 | 0,542 % | 0,973 |
| CMF | Bancario + Externo | Sky / No Sky | $7.782.971 | 0,174 % | 0,955 |
| No se utilizan | Bancario + Externo | Sky / No Sky | $7.782.971 | 0,174 % | 0,955 |
| Bancario | Bancario + Externo | Sky / No Sky | $7.749.706 | 0,583 % | 0,971 |
| Bancario | Bancario + CMF + Externo | Sky / No Sky | $7.736.872 | 0,471 % | 0,971 |
| No se utilizan | Bancario + CMF + Externo | 1 / 3 / 6 cuotas | $7.672.909 | 0,551 % | 0,966 |
| Bancario | Bancario + CMF + Externo | 1 / 3 / 6 cuotas | $7.543.387 | 0,573 % | 0,969 |
| CMF | Bancario + CMF + Externo | 1 / 3 / 6 cuotas | $7.491.063 | 0,847 % | 0,968 |
| Bancario y CMF | Bancario + Externo | Sky / No Sky | $7.464.741 | 0,024 % | 0,974 |
| CMF | Bancario + CMF | 1 / 3 / 6 cuotas | $7.454.074 | 0,852 % | 0,965 |
| Bancario y CMF | Bancario + CMF | Sky / No Sky | $7.281.447 | 0,702 % | 0,948 |
| Bancario y CMF | Bancario + Externo | 1 / 3 / 6 cuotas | $7.233.088 | 1,339 % | 0,965 |
| Bancario | Bancario + Externo | 1 / 3 / 6 cuotas | $7.233.088 | 1,339 % | 0,965 |
| CMF | Bancario + Externo | 1 / 3 / 6 cuotas | $7.228.929 | 1,324 % | 0,962 |
| No se utilizan | Bancario + Externo | 1 / 3 / 6 cuotas | $7.228.929 | 1,324 % | 0,962 |
| CMF | CMF + Externo | Sky / No Sky | $7.149.173 | 0,954 % | 0,922 |
| Bancario y CMF | Bancario + CMF | 1 / 3 / 6 cuotas | $7.098.034 | 1,146 % | 0,953 |
| Bancario y CMF | Bancario + CMF + Externo | 1 / 3 / 6 cuotas | $7.067.411 | 1,208 % | 0,963 |
| Bancario y CMF | CMF + Externo | Sky / No Sky | $7.039.349 | 0,962 % | 0,917 |
| CMF | Bancario + CMF + Externo | Sky / No Sky | $6.954.398 | 1,134 % | 0,946 |
| No se utilizan | Bancario + CMF + Externo | Sky / No Sky | $6.937.706 | 1,276 % | 0,910 |
| Bancario | Bancario + CMF | 1 / 3 / 6 cuotas | $6.619.180 | 0,585 % | 0,945 |
| No se utilizan | Bancario + CMF | Sky / No Sky | $6.590.409 | 1,952 % | 0,907 |
| Bancario y CMF | Bancario | 1 / 3 / 6 cuotas | $6.463.466 | 0,092 % | 0,905 |
| Bancario | Bancario | 1 / 3 / 6 cuotas | $6.463.466 | 0,092 % | 0,905 |
| CMF | Bancario + CMF | Sky / No Sky | $6.426.057 | 2,316 % | 0,916 |
| No se utilizan | Bancario + CMF | 1 / 3 / 6 cuotas | $6.345.601 | 1,643 % | 0,940 |
| Bancario y CMF | Bancario | Sky / No Sky | $6.196.920 | 0,027 % | 0,933 |
| Bancario | Bancario | Sky / No Sky | $6.196.920 | 0,027 % | 0,933 |
| Bancario y CMF | CMF + Externo | 1 / 3 / 6 cuotas | $5.999.848 | 2,467 % | 0,891 |
| CMF | CMF + Externo | 1 / 3 / 6 cuotas | $5.999.848 | 2,467 % | 0,891 |
| Bancario | Bancario + CMF | 50 / 50 | $5.799.258 | 0,186 % | 0,987 |
| No se utilizan | Bancario + CMF | 50 / 50 | $5.765.372 | 0,301 % | 0,978 |
| Bancario | Bancario + CMF + Externo | 50 / 50 | $5.739.933 | 0,055 % | 0,984 |
| Bancario y CMF | Bancario + Externo | 50 / 50 | $5.636.793 | 0,172 % | 0,968 |
| Bancario | Bancario + Externo | 50 / 50 | $5.636.793 | 0,172 % | 0,968 |
| CMF | Bancario + Externo | 50 / 50 | $5.629.979 | 0,171 % | 0,968 |
| No se utilizan | Bancario + Externo | 50 / 50 | $5.629.979 | 0,171 % | 0,968 |
| No se utilizan | Bancario + CMF + Externo | 50 / 50 | $5.595.533 | 0,266 % | 0,977 |
| CMF | Bancario | 1 / 3 / 6 cuotas | $5.579.396 | 0,100 % | 0,884 |
| No se utilizan | Bancario | 1 / 3 / 6 cuotas | $5.579.396 | 0,100 % | 0,884 |
| CMF | Bancario | SKY / No SKY | $5.493.176 | 0,273 % | 0,900 |
| No se utilizan | Bancario | SKY / No SKY | $5.493.176 | 0,273 % | 0,826 |
| CMF | Bancario + CMF + Externo | 50 / 50 | $5.486.030 | 0,311 % | 0,946 |
| Bancario y CMF | Bancario + CMF + Externo | 50 / 50 | $5.464.895 | 0,699 % | 0,968 |
| Bancario y CMF | Bancario + CMF | 50 / 50 | $5.424.643 | 0,762 % | 0,966 |
| CMF | Bancario + CMF | 50 / 50 | $5.409.428 | 0,859 % | 0,954 |
| No se utilizan | Bancario + CMF | SKY / No SKY | $5.277.602 | 3,636 % | 0,902 |

Tabla D.3: Comparación resultados uso de variables externas, segmentaciones y modelos ordenados por utilidad de mayor a menor. Parte 1. Fuente: *Elaboración propia*

| Uso variables externas | Modelo | Segmentación | Utilidad | No pago | AUC |
|---|---|---|---|---|---|
| No se utilizan | CMF + Externo | SKY / No SKY | $5.185.057 | 3,666 % | 0,878 |
| No se utilizan | CMF + Externo | SKY / No SKY | $5.185.057 | 3,666 % | 0,897 |
| No se utilizan | CMF + Externo | 70 / 30 | $5.179.009 | 0,791 % | 0,939 |
| No se utilizan | CMF + Externo | 70 / 30 | $5.179.009 | 0,791 % | 0,939 |
| No se utilizan | CMF + Externo | 1 / 3 / 6 cuotas | $5.022.581 | 3,958 % | 0,831 |
| No se utilizan | CMF + Externo | 1 / 3 / 6 cuotas | $5.022.581 | 3,958 % | 0,831 |
| CMF | Bancario + CMF + Externo | 70 / 30 | $4.917.218 | 0,671 % | 0,961 |
| Bancario y CMF | Bancario | 50 / 50 | $4.836.758 | 0,060 % | 0,947 |
| No se utilizan | Bancario | 50 / 50 | $4.836.758 | 0,060 % | 0,947 |
| CMF | Bancario | 50 / 50 | $4.809.438 | 0,253 % | 0,943 |
| No se utilizan | Bancario | 50 / 50 | $4.809.438 | 0,253 % | 0,943 |
| No se utilizan | Bancario + CMF + Externo | 70 / 30 | $4.807.460 | 1,405 % | 0,950 |
| Bancario y CMF | Bancario + CMF + Externo | 70 / 30 | $4.721.607 | 1,274 % | 0,944 |
| CMF | Bancario + Externo | 70 / 30 | $4.595.159 | 2,021 % | 0,938 |
| No se utilizan | Bancario + Externo | 70 / 30 | $4.595.159 | 2,021 % | 0,938 |
| Bancario y CMF | CMF | 50 / 50 | $4.253.517 | 0,453 % | 0,871 |
| CMF | CMF | 50 / 50 | $4.253.517 | 0,453 % | 0,871 |
| Bancario y CMF | Bancario + CMF | 70 / 30 | $4.229.530 | 2,038 % | 0,925 |
| No se utilizan | Bancario + CMF | 70 / 30 | $4.219.285 | 2,147 % | 0,932 |
| No se utilizan | Bancario + CMF | 70 / 30 | $4.160.643 | 2,421 % | 0,923 |
| Bancario y CMF | CMF + Externo | 70 / 30 | $4.008.997 | 2,801 % | 0,878 |
| CMF | CMF + Externo | 70 / 30 | $4.008.997 | 2,801 % | 0,878 |
| Bancario y CMF | CMF | SKY / No SKY | $3.691.981 | 0,225 % | 0,792 |
| CMF | CMF | SKY / No SKY | $3.691.981 | 0,225 % | 0,792 |
| No se utilizan | CMF + Externo | 50 / 50 | $3.652.005 | 2,923 % | 0,837 |
| No se utilizan | CMF + Externo | 50 / 50 | $3.652.005 | 2,923 % | 0,837 |
| Bancario y CMF | Externo + Bancario | 70 / 30 | $3.615.148 | 3,448 % | 0,897 |
| Bancario | Externo + Bancario | 70 / 30 | $3.615.148 | 3,448 % | 0,897 |
| Bancario | CMF | 70 / 30 | $3.504.803 | 0,115 % | 0,829 |
| No se utilizan | CMF | 70 / 30 | $3.504.803 | 0,115 % | 0,829 |
| CMF | Bancario + CMF | 70 / 30 | $3.368.019 | 3,384 % | 0,894 |
| CMF | Bancario | 70 / 30 | $3.323.173 | 0,480 % | 0,870 |
| No se utilizan | Bancario | 70 / 30 | $3.323.173 | 0,480 % | 0,870 |
| Bancario y CMF | Bancario | 70 / 30 | $3.304.152 | 0,758 % | 0,874 |
| Bancario | Bancario | 70 / 30 | $3.304.152 | 0,758 % | 0,874 |
| Bancario y CMF | CMF + Externo | 50 / 50 | $3.071.311 | 4,290 % | 0,819 |
| CMF | CMF + Externo | 50 / 50 | $3.071.311 | 4,290 % | 0,819 |
| No se utilizan | Bancario + CMF + Externo | 70 / 30 | $2.895.197 | 4,539 % | 0,884 |
| Bancario y CMF | CMF | 70 / 30 | $2.361.331 | 0,000 % | 0,769 |
| CMF | CMF | 70 / 30 | $2.361.331 | 0,000 % | 0,769 |
| Bancario y CMF | Externo | 50 / 50 | $2.095.365 | 5,694 % | 0,732 |
| Bancario | Externo | 50 / 50 | $2.095.365 | 5,694 % | 0,732 |
| CMF | Externo | 50 / 50 | $2.095.365 | 5,694 % | 0,732 |
| No se utilizan | Externo | 50 / 50 | $2.095.365 | 5,694 % | 0,732 |
| Bancario y CMF | Externo | 70 / 30 | $2.095.365 | 5,694 % | 0,743 |
| Bancario | Externo | 70 / 30 | $2.095.365 | 5,694 % | 0,743 |
| CMF | Externo | 70 / 30 | $2.095.365 | 5,694 % | 0,743 |
| No se utilizan | Externo | 70 / 30 | $2.095.365 | 5,694 % | 0,743 |

Tabla D.4: Comparación resultados uso de variables externas, segmentaciones y modelos ordenados por utilidad de mayor a menor. Parte 2. Fuente: *Elaboración propia*

| Uso variables externas | Modelo | Segmentación | Utilidad | No pago | AUC |
|---|---|---|---|---|---|
| Bancario | CMF | SKY / No SKY | $1.709.518 | 0,140 % | 0,727 |
| No se utilizan | CMF | SKY / No SKY | $1.709.518 | 0,140 % | 0,700 |
| Bancario y CMF | Externo | 1 / 3 / 6 cuotas | $1.671.719 | 7,738 % | 0,734 |
| Bancario | Externo | 1 / 3 / 6 cuotas | $1.671.719 | 7,738 % | 0,734 |
| CMF | Externo | 1 / 3 / 6 cuotas | $1.671.719 | 7,738 % | 0,734 |
| No se utilizan | Externo | 1 / 3 / 6 cuotas | $1.671.719 | 7,738 % | 0,734 |
| Bancario y CMF | CMF | 1 / 3 / 6 cuotas | $1.278.389 | 0,148 % | 0,658 |
| CMF | CMF | 1 / 3 / 6 cuotas | $1.278.389 | 0,148 % | 0,658 |
| Bancario | CMF | 1 / 3 / 6 cuotas | $1.260.307 | 0,000 % | 0,669 |
| No se utilizan | CMF | 1 / 3 / 6 cuotas | $1.260.307 | 0,000 % | 0,669 |
| Bancario | CMF | 50 / 50 | $1.079.227 | 0,791 % | 0,766 |
| No se utilizan | CMF | 50 / 50 | $1.079.227 | 0,791 % | 0,766 |
| Bancario y CMF | Externo | SKY / No SKY | $569.005 | 8,901 % | 0,745 |
| Bancario | Externo | SKY / No SKY | $569.005 | 8,901 % | 0,745 |
| CMF | Externo | SKY / No SKY | $569.005 | 8,901 % | 0,745 |
| No se utilizan | Externo | SKY / No SKY | $569.005 | 8,901 % | 0,810 |

Tabla D.5: Comparación resultados uso de variables externas, segmentaciones y modelos ordenados por utilidad de mayor a menor. Parte 3. Fuente: *Elaboración propia*

## D.7. Propuesta de implementación y monitoreo

Los pasos recomendados para implementar el modelo en producción, de manera general, son los siguientes:

- **Creación de reglas fuera del modelo**: El primer paso, una vez establecido el nuevo modelo, consiste en implementar restricciones adicionales que se relacionen directamente con los usuarios y las probabilidades generadas por el modelo. Esto podría considerarse como una segunda segmentación, que ajusta los puntos de corte según los perfiles de los usuarios. Por ejemplo, se podría permitir que aquellos usuarios que hayan pagado más de un determinado porcentaje de sus compras con Cleo puedan presentar un riesgo mayor que otros.
  - **Responsables de la etapa**: Equipo de BNPL
  - **Posibles desafíos**: Encontrar valores óptimos para la segmentación de perfiles puede resultar complejo, especialmente si el tiempo disponible es limitado.
- **Implementación web para pruebas**: El siguiente paso es poner en funcionamiento el modelo en un entorno web para realizar pruebas y verificar el correcto funcionamiento de todos los componentes necesarios.
  - **Responsables de la etapa**: Equipo de desarrollo
  - **Posibles desafíos**: Puede haber problemas de compatibilidad entre el lenguaje utilizado por el equipo de desarrollo y el empleado durante el desarrollo del modelo, lo que requeriría mantener una comunicación constante para realizar posibles modificaciones.

- **Implementación web en producción**: Después de confirmar que el modelo funciona correctamente en el entorno de pruebas, se procede a su implementación en producción para utilizarlo en transacciones reales.
  - **Responsables de la etapa**: Equipo de desarrollo
  - **Posibles desafíos**: No se anticipan grandes desafíos en esta fase, ya que los problemas deberían haberse abordado en la etapa anterior.
- **Monitoreo y actualización**: Una vez implementado, es fundamental mantener un monitoreo constante de los resultados del modelo. Entre las métricas útiles se incluyen el porcentaje de compras aceptadas, las probabilidades generadas y el grado de morosidad resultante. Se recomienda realizar el monitoreo cada cuatro semanas para acumular una cantidad suficiente de datos, y considerar una actualización del modelo cada tres meses para incorporar los datos más recientes, tal como se ilustra en la Figura D.12. Esta etapa consideraría también un gasto adicional debido a la persona encargada.
  - **Responsables de la etapa**: Cientista de datos
  - **Posibles desafíos**: El volumen de transacciones podría presentar problemas: si es bajo, no se podrían generalizar correctamente las conclusiones del análisis, lo que requeriría extender el tiempo de monitoreo; si es alto, el monitoreo debería realizarse con mayor frecuencia.

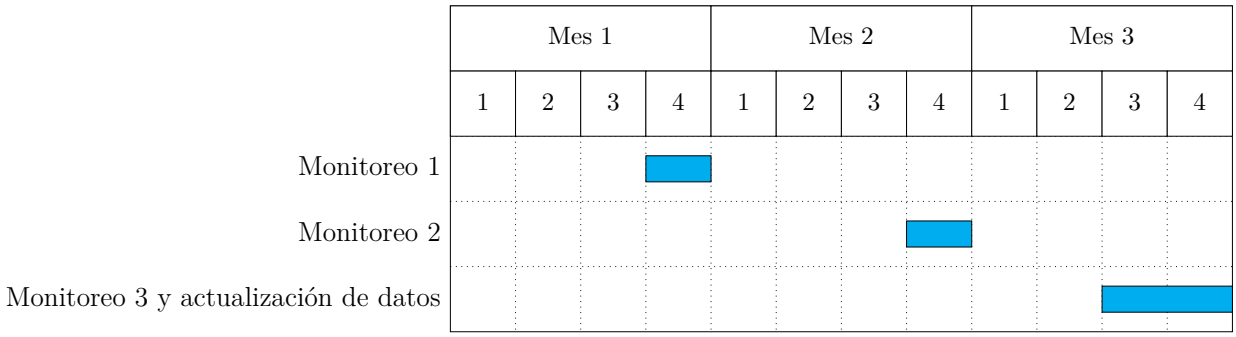


Figura D.12: Propuesta de intervalos de revisión del modelo en carta Gantt.
Fuente: *Elaboración propia*

## D.8. Resumen detallado de hallazgos en el desarrollo

En esta sección se presentan de manera detallada los hallazgos de la Sección 7, referente al desarrollo del proyecto. También se ilustran las relaciones entre los objetivos específicos y las fases de la metodología en la Tabla D.6, junto con los principales hallazgos encontrados.

| Objetivo | Fases en que se desarrolló | Principales hallazgos |
|---|---|---|
| **Analizar datos de los usuarios de BNPL** | • Comprensión del negocio<br>• Comprensión de los datos | • Utilidad positiva bajo el 8 % de no pago<br>• Buenos y malos pagadores presentan diferencias |
| **Revisión de la literatura** | • Comprensión del negocio<br>• Comprensión de los datos<br>• Preparación de los datos<br>• Modelado<br>• Evaluación<br>• Despliegue | • Mejores algoritmos para utilizar: Bosques aleatorios, XGBoost y LightGBM<br>• Métricas adecuadas al contexto de negocio<br>• Librerías útiles para el desarrollo del ttrabajo |
| **Diseño y desarrollo de modelos** | • Preparación de los datos<br>• Modelado | • Creación de variables<br>• Técnica de ensamblaje |
| **Preparación y validación** | • Evaluación<br>• Despliegue | • Diferencia entre modelos complejos y simples<br>• Revisión de sesgo<br>• Aprendizaje de empaquetado |

Tabla D.6: Tabla resumen de objetivos en el desarrollo. Fuente: *Elaboración propia*

Así, se tiene lo siguiente para cada objetivo específico:

- **Analizar datos de los usuarios de BNPL:** Abordado mediante las fases de comprensión del negocio y de los datos. Este objetivo se completó al investigar las diferencias entre medias de grupos de buenos y malos pagadores, lo que concluyó en una confirmación de que las variables que se poseían en las bases de datos contribuían a identificar los tipos de usuarios.

  Además, las reuniones con los equipos y la investigación de los datos permitieron establecer un límite de tasa de no pago inferior al 8 % para asegurar la rentabilidad del servicio.

- **Revisión de la literatura:** Este objetivo se desarrolló a medida que el proyecto avanzaba, por lo que se ubica de forma transversal en las fases de la metodología. Se investigaron los algoritmos a utilizar, se buscó documentación sobre librerías para su personalización y se definieron los métodos de evaluación.

En pos del tiempo utilizado y de las fuentes revisadas, el objetivo se percibe como completo, resultando en la utilización de las mejores prácticas sobre algoritmos y modelamiento en el contexto del riesgo crediticio con los recursos disponibles.

- **Diseño y desarrollo de modelos:** Este objetivo requirió la mayor parte del tiempo de trabajo debido a que su realización dependía de las fases de preparación de datos y modelado, las cuales, sumado a las iteraciones realizadas y a la necesidad de utilizar una base de datos óptima, utilizaron más tiempo del presupuestado, ajustando el plazo de las últimas fases.

  A pesar de las complicaciones en términos de tiempo, este objetivo se cumplió, ya que se obtuvo un conjunto de variables con gran calidad de información y con una reducción significativa del ruido en comparación a la base inicial, siendo el ruido datos o alteraciones que bajan la calidad de las predicciones (Gupta y Gupta, 2019). También se aprendió sobre la utilización de la técnica de ensamblaje, que combina las predicciones de distintos modelos para reducir la varianza y el error de los modelos individuales.

- **Preparación y validación:** Objetivo desarrollado a partir de la evaluación de los modelos y el despliegue final. Se realizaron evaluaciones tanto cuantitativas como cualitativas, considerando posibles actualizaciones futuras para el modelo de negocio. Este objetivo se considera completo, ya que se determinaron los mejores modelos con base en las métricas propuestas, se validaron con el equipo, se evaluó su nivel de sesgo y finalmente se entregó a la compañía el modelo óptimo escogido.

  Durante el desarrollo del objetivo, se observó la diferencia en los resultados entre los modelos simples y los modelos más complejos, como los de ensamblaje. Además, al comparar esta mayor complejidad con el uso de variables externas, se abrió la discusión sobre la posibilidad de mejorar las predicciones incorporando otras variables adicionales.

  Además, se aprendió sobre variables protegidas en la revisión del sesgo de los modelos y sobre el método de ejecución y de entrega de datos para la implementación de un modelo en producción, sumado a la creación de su correspondiente documentación y plan de seguimiento.

En síntesis, los objetivos fueron completados al finalizar las fases de la metodología correspondientes. Se lograron hallazgos importantes tanto para el proyecto como para la compañía y para futuros proyectos. Además, se obtuvo un aprendizaje significativo para el estudiante en el ámbito de la ciencia de datos.