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

DESARROLLO DE UN MODELO DE MACHINE LEARNING DE ESTIMACIÓN DE INGRESOS PARA EL OTORGAMIENTO DE CRÉDITO EN INSTITUCIONES FINANCIERAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

OCTAVIO ADOLFO LÓPEZ OLEA

PROFESOR GUÍA:
HUGO SÁNCHEZ RAMÍREZ

MIEMBROS DE LA COMISIÓN:
RONALD FISCHER BARKAN
CARLOS PULGAR ARATA

SANTIAGO DE CHILE
2024

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

POR: OCTAVIO ADOLFO LÓPEZ OLEA

FECHA: 2024

PROF. GUÍA: HUGO SÁNCHEZ RAMÍREZ

DESARROLLO DE UN MODELO DE MACHINE LEARNING DE ESTIMACIÓN DE INGRESOS PARA EL OTORGAMIENTO DE CRÉDITO EN INSTITUCIONES FINANCIERAS

La industria del retail financiero ha experimentado un crecimiento notable en Chile durante las últimas dos décadas, consolidándose como un actor clave en la emisión de tarjetas de crédito y la oferta de servicios financieros. Este crecimiento ha venido acompañado de desafíos significativos, especialmente en la gestión del riesgo crediticio, exacerbados por el aumento de la morosidad en un contexto económico desafiante post-pandemia.

En respuesta a estos desafíos, este proyecto se centró en el desarrollo de un modelo de Machine Learning para la estimación de ingresos, dirigido al negocio financiero de una de las principales cadenas farmacéuticas del país. El propósito del proyecto fue mejorar la precisión y confiabilidad de las estimaciones de ingresos de los clientes y potenciales clientes, buscando optimizar la asignación de cupos para tarjetas de crédito y, así, disminuir el riesgo de default y las tasas de morosidad.

Como parte del proyecto, el rol del memorista incluyó la recopilación y preprocesamiento de datos, el desarrollo y validación del modelo, y la comparación de su rendimiento con las estimaciones externas. La metodología empleada abarcó técnicas avanzadas de análisis de datos y modelado predictivo, utilizando un enfoque de aprendizaje automático para mejorar la precisión de las predicciones.

Los resultados obtenidos fueron significativos: el modelo desarrollado redujo en más de un 25% el Error Absoluto Medio (MAE) en comparación con las estimaciones proporcionadas por el proveedor externo. Esta mejora no solo optimiza la precisión de las decisiones crediticias, sino que también establece una base sólida para el desarrollo de futuras herramientas de gestión de riesgos internas dentro de la institución.

☺ .

Agradecimientos

Quisiera comenzar expresando mi más sincero agradecimiento al Profesor Guía de esta memoria, Hugo Sánchez, por brindarme su tiempo y conocimiento, elementos fundamentales para el desarrollo de este proyecto. Asimismo, quiero agradecer al Profesor Carlos Pulgar por su excelente disposición y amabilidad. También extiendo mi gratitud a cada profesor que ha sido parte de mi formación académica en la Universidad de Chile, demostrando siempre la importancia de ser personas íntegras.

A mis amigos que conocí durante mi estancia por la Universidad, muchas gracias, pude encontrar un grupo de personas increíbles con quienes reír, les deseo lo mejor en todo lo que viene por delante.

A todos mis gatos y perros, a los que están y ya no, llevo a cada uno de ustedes dentro de mí.

A Josefina, por ser la compañera que siempre anhelé, por hacerme feliz e impulsarme a superarme cada día, mi más profundo agradecimiento.

Por último, pero no menos importante, quisiera agradecer a mi familia: Juan Carlos, Fabiola, Paulina y Carlos. Les agradezco su apoyo incondicional y la formación que me han brindado a lo largo de mi vida, hay una parte de ustedes en cada logro.

Tabla de Contenido

Capítulo 1 : Introducción *1*

1.1. Antecedentes Generales **1**

1.2. Descripción del Problema **3**

1.3. Descripción y Justificación del Proyecto **4**

1.3.1. Tarjeta Abierta 4

1.3.2. Dependencia a Proveedores Externos 4

1.3.3. Informe de Deuda Consolidada 5

Capítulo 2 : Objetivos *6*

2.1. Objetivo General **6**

2.2. Objetivos Específicos **6**

2.3. Alcances **6**

Capítulo 3 : Marco Conceptual *7*

3.1. Machine Learning **7**

3.2. Modelos Supervisados **7**

3.3. Modelos de Regresión **7**

3.1. Gradient Boosting Regressor **8**

3.2. Hiperparámetros **9**

3.3. Validación Cruzada **10**

3.4. Métricas de Rendimiento **10**

3.4.1. Mean Absolute Error (MAE) 10

3.4.2. Median Absolute Error (MDAE) 10

3.4.3. Coeficiente de Determinación (R^2) 11

3.5. Transformación Logarítmica 11

3.6. Z-score 11

Capítulo 4 : Metodología 12

4.1. Caracterización de la gestión de créditos 12

4.2. Recopilación de Data 12

4.3. Set de Ingresos y Tratamiento de la Data 12

4.4. Pre Procesamiento de Características 12

4.5. Análisis de Características 13

4.6. Cálculo de Rendimiento Estimaciones Externas 13

4.7. Desarrollo del Modelo 13

4.8. Test 13

Capítulo 5 : Desarrollo y Resultados 14

5.1. Caracterización de la gestión de créditos 14

5.1.1. Requisitos Generales Productos Financieros 14

5.1.2. Base Aprobada para otorgamiento de Crédito 14

5.1.3. Asignación de Cupos y Monto de Crédito 14

5.2. Recopilación de Data 15

5.2.1. Dataset Ingresos Comprobados 15

5.2.2. Dataset Variables Financieras 15

5.2.3. Dataset Variables Demográficas 15

5.2.4. Dataset Variables Retail 16

5.3. Set de Ingresos y Tratamiento de la data 17

5.3.1. Clientes con más de una Acreditación de Ingresos 17

5.3.2. Actualización de Ingresos 17

5.3.3. Métricas y Distribución del Set de Ingresos 18

5.4. Pre-procesamiento de Características 20

5.4.1. Características Numéricas 20

5.4.2. Características Categóricas 20

5.5. Análisis de Características 21

5.5.1. Características Financieras 21

5.5.2. Características Demográficas 24

5.5.3. Características Retail 26

5.6. Cálculo Rendimiento Estimaciones Externas 28

5.7. Desarrollo del Modelo 29

5.7.1. Data Usage 29

5.7.2. Segmentación 30

5.7.3. Entrenamiento 30

5.8. Test 33

Capítulo 6 : Discusiones 34

6.1. Limitaciones del Modelo y los Datos 34

6.2. Desarrollo de Herramientas Internas y Expansión hacia la Gestión Integral del Riesgo Crediticio 34

6.3. Propuestas Estratégicas a Futuro 35

Capítulo 7 : Conclusiones 36

Bibliografía 37

Índice de Tablas

Tabla 1: Factores de Conversión de Ingresos. 17
Tabla 2: Métricas Set final de Ingresos Comprobados. 18
Tabla 3: Métricas de Rendimiento de Estimaciones de Ingresos Externas para el Año 2024. 28
Tabla 4: Métricas de Rendimiento Estimaciones de Ingresos Externas para el Año 2020 a 2023. 28
Tabla 5: Cantidad de Registros para Sets 30
Tabla 6: Valores óptimos Hiperparámetros gbm. 31
Tabla 7: Resultados Modelos en Entrenamiento. 31
Tabla 8: Resultados Modelos en Test. 33

Índice de Figuras

Figura 1: Distribución de colocaciones financieras en el sistema total, consumo + vivienda, ... 2
Figura 2: Evolución Mensual porcentajes de morosidad en colocaciones asociadas y 3
Figura 3: Ilustración del proceso de Gradient Boosting. 9
Figura 4: Distribución Set de Ingreso Final. 18
Figura 5: Distribución Set de Ingresos en escala Logarítimica. 19
Figura 6: Distribución variable CRE_CONSUMOS en Set de Ingresos Comprobados. 21
Figura 7: Distribución variable CRE_CONSUMOS población total en Sistema financiero. 22
Figura 8: Distribución variable CRE_CONSUMOS. 22
Figura 9: Boxplot variable CRE_CONSUMOS por Quintil de ingresos 23
Figura 10:Tendencia de Ingreso por Edad 24
Figura 11: Boxplot variable Género 25
Figura 12: Boxplot variable Pensionado 25
Figura 13: Gráfico de caja variable GSE 26
Figura 14: Boxplot variable Necesidades 27
Figura 15: Gráfico de dispersión de errores para el Set de Ingresos del año 2024 29
Figura 16: Curva de Aprendizaje Modelo Financiero. 32
Figura 17: Boxplots de errores para Ingresos sobre y bajo 1 millón de pesos Modelo Financiero. 33

Capítulo 1: Introducción

1.1. Antecedentes Generales

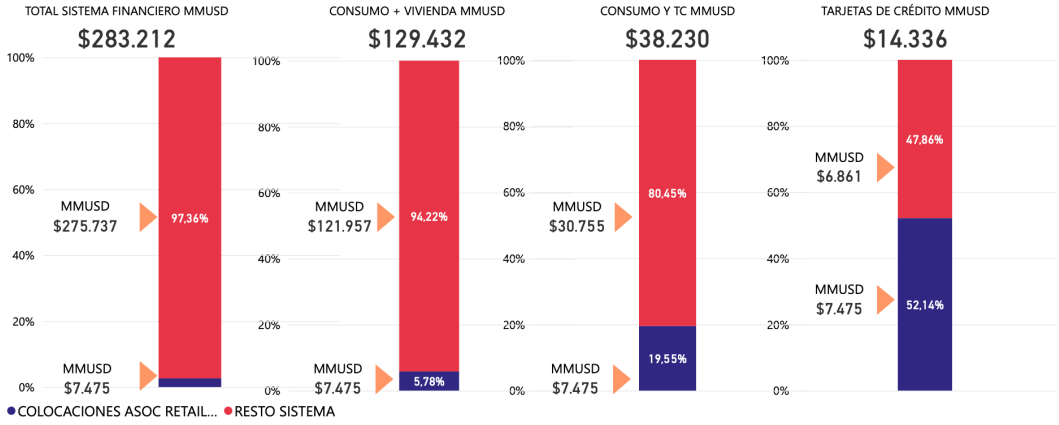
La industria del retail, también conocida como la industria minorista, es uno de los sectores económicos de mayor importancia en la economía global, cuyo objetivo es la venta de bienes y servicios al consumidor final, actuando como un puente entre los productores y los consumidores, facilitando la distribución en el mercado. En el ámbito nacional, las ventas del retail representan aproximadamente el 20 % del PIB, con un movimiento de alrededor de 63.600 millones de dólares en el año 2022 (Diario Financiero Chile, 2023), generando más de 1,5 millones de puestos de trabajo (Ortiz, 2019).

Otra de las principales industrias en la economía, es el sector financiero, que abarca una amplia gama de actividades, desde la banca hasta los seguros y la inversión. Respecto al panorama nacional, el sistema bancario chileno se caracteriza por tener una gran estabilidad, posicionándose como el más estable de la región, destacándose en aspectos como la capitalización de mercado, participación en el crédito doméstico y número de sucursales por habitantes. Este desarrollo del sistema financiero chileno ha facilitado un mayor acceso a servicios y productos financieros. En particular, el uso de cuentas corrientes, líneas de crédito, tarjetas de crédito y débito supera la mediana de los países comparados en el estudio realizado por el Banco Central de Chile (Banco Central de Chile, 2021).

La constante búsqueda de innovación y transformación de ambas industrias durante las últimas décadas, ha desarrollado el fenómeno del retail financiero, que consiste en la oferta de productos y servicios financieros, principalmente tarjetas de crédito y créditos de consumo, por parte de marcas del sector minorista. Al expandir el negocio, las empresas del retail logran incrementar y diversificar sus fuentes de ingreso utilizando el posicionamiento que ya posee la marca. A la vez, la oferta de servicios y productos financieros ayuda a la captación de nuevos clientes como a la retención de los clientes existentes, mediante el otorgamiento de crédito y promociones exclusivas. En conjunto, la expansión del negocio puede generar una mayor presencia de la marca en el mercado, además de economías de escala.

Como se puede ver en la Figura 1, en el panorama nacional las entidades financieras asociadas al retail representan aproximadamente un 2,5% del total de stock de colocaciones del sistema financiero, mientras que lo que respecta al stock de colocaciones de tarjetas de créditos, representan un 52,14% del total con $7.775 MM USD (Asociación Retail Financiero, 2023), demostrando la importancia de este sector en la economía nacional, principalmente como medio de compra de bienes y servicios en el comercio.

Figura 1: Distribución de colocaciones financieras en el sistema total, consumo + vivienda, consumo y tarjetas de crédito.



Nota. De Compendio Estadístico Mensual de la Industria del Crédito, por Asociación Retail Financiero [ARF], noviembre de 2023. Recuperado de https://retailfinanciero.org/wp-content/uploads/2024/03/Reporte-CMF-Bcentral-Sep-2023_vnov.pdf.

En este contexto, una de las principales cadenas farmacéuticas del país, que cuenta con una significativa participación de mercado y numerosos locales y colaboradores, mediante su filial dedicada al negocio financiero, ofrece productos y servicios de este tipo. Aprovechando el posicionamiento de la marca y una amplia red de sucursales, logra llegar a los consumidores de manera masiva, brindándoles acceso a crédito.

Por motivos de confidencialidad, el sponsor del proyecto será nombrado genéricamente en el transcurso de este documento como Negocio Financiero.

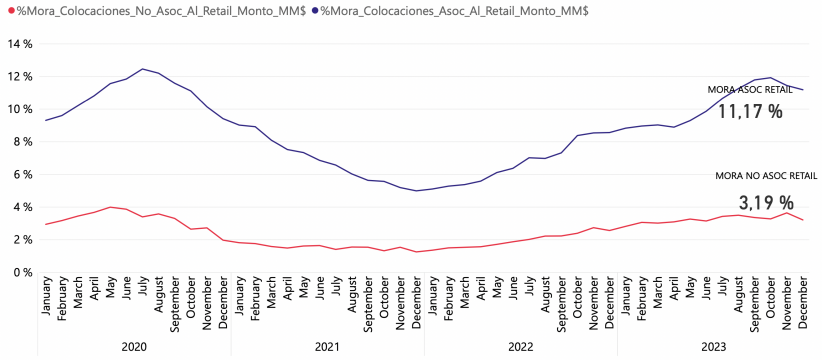
Los productos que ofrece Negocio Financiero son créditos de consumo y una tarjeta de crédito diseñada para clientes de la farmacia, representando una herramienta de captación y fidelización al ofrecer beneficios exclusivos y descuentos en productos y servicios. Esta integración vertical le permite diversificar sus fuentes de ingresos y crear sinergias entre sus diferentes líneas de negocio, otorgándole una ventaja competitiva significativa en el mercado.

1.2. Descripción del Problema

El riesgo crediticio se refiere a la posibilidad de que una entidad financiera enfrente pérdidas debido a que los deudores no cumplan con el pago de sus obligaciones. Este riesgo puede surgir de la insolvencia del deudor, que puede ser causada por condiciones económicas adversas o por una gestión inadecuada del crédito por parte de la entidad. El default es una manifestación específica de este riesgo, donde un prestatario no cumple con sus obligaciones de pago según lo acordado. Esto puede presentarse de diversas formas, como la falta de pago de cuotas, retrasos en los pagos o la incapacidad total de pagar el préstamo.

En el contexto nacional, se ha observado un aumento en las tasas de morosidad en tarjetas de crédito. Este incremento se ha atribuido a factores económicos adversos, como la desaceleración económica y el aumento del desempleo, que han afectado la capacidad de pago de los consumidores (CIEDESS, 2023). La Figura 2 ilustra la evolución mensual de las tasas de morosidad en tarjetas de crédito, destacando esta tendencia creciente en los últimos años.

Figura 2: Evolución Mensual porcentajes de morosidad en colocaciones asociadas y no asociadas al retail.



Nota. De Compendio Estadístico Mensual de la Industria del Crédito, por Asociación Retail Financiero [ARF], noviembre de 2023. Recuperado de https://retailfinanciero.org/wp-content/uploads/2024/03/Reporte-CMF-Bcentral-Sep-2023_vnov.pdf.

Negocio Financiero no ha estado exento de esta problemática, enfrentando un aumento significativo en sus gastos asociados a la gestión del riesgo crediticio debido a un incremento en las tasas de morosidad. Esta problemática ha ejercido una presión financiera adicional sobre la empresa, obligándola a destinar más recursos a la evaluación del riesgo crediticio, la recuperación de deudas incobrables y otras actividades relacionadas. Específicamente, los gastos destinados a mitigar el riesgo se incrementaron en aproximadamente un 100% durante el último año.

Uno de los problemas identificados en la gestión del riesgo crediticio dentro de la empresa, es la utilización de una estimación de ingresos individuales proporcionada por una empresa externa, la cual ha demostrado ser inexacta y poco confiable. Persistir en el uso de esta herramienta conlleva varias consecuencias negativas que impactan directamente en la rentabilidad de la empresa.

En primer lugar, basar las decisiones de otorgamiento de crédito en una estimación poco confiable, expone a Negocio Financiero el riesgo de conceder créditos a personas cuya capacidad de pago

real no coincide con la información proporcionada. Esto puede resultar en una mayor tasa de incumplimiento de pago y, consecuentemente, en pérdidas financieras para la empresa.

Además, el incremento en los problemas de cobranza derivados de la falta de precisión en la estimación de ingresos, genera una carga adicional en los procesos administrativos y de gestión de créditos, aumentando los costos operativos y reduciendo la eficiencia operativa. Este impacto negativo se refleja en la rentabilidad general de la empresa y su capacidad para competir en un mercado cada vez más exigente.

Reconocer y abordar adecuadamente esta problemática permite a Negocio Financiero sentar las bases para reducir significativamente su exposición al riesgo crediticio. Al desarrollar e implementar una metodología interna más sólida para evaluar la solvencia de los prestatarios, la empresa puede mejorar la precisión de sus decisiones de crédito y reducir las tasas de morosidad.

1.3. Descripción y Justificación del Proyecto

1.3.1. Tarjeta Abierta

Durante el transcurso del año en que se realizó este proyecto, Negocio Financiero cambiará el enfoque de la tarjeta de crédito, pasando de ser una tarjeta cerrada, es decir que admite compras solo dentro de la farmacia de la compañía, a ser una tarjeta abierta Visa, que admite compras en todo el comercio. Este cambio se realiza con el propósito de hacer al producto más atractivo para los actuales y nuevos clientes, entregándoles más flexibilidad y libertad de uso del crédito otorgado, lo que posiblemente generará un mayor uso de la tarjeta y además un aumento en el número de clientes.

El principal desafío para las entidades financieras al ofrecer una tarjeta abierta es la mayor propensión al riesgo que conlleva. Al permitir el uso de la tarjeta a todo el comercio, se incrementa significativamente el riesgo asociado con el crédito. Más precisamente, el aumento en la libertad de uso puede llevar a un incremento en el endeudamiento de los clientes, quienes podrían gastar más allá de sus capacidades de pago. Esto no solo incrementa el riesgo de impagos y morosidad, sino que también puede llevar a una mayor exposición al fraude y al uso indebido de la tarjeta (Zack & Newton, 2020; Visa, 2022). Por lo tanto, Negocio Financiero debe implementar herramientas de gestión de riesgo más sofisticadas para mitigar estos posibles inconvenientes.

1.3.2. Dependencia a Proveedores Externos

Actualmente, las principales herramientas de gestión crediticia utilizadas en Negocio Financiero, como la estimación de ingresos y el score de riesgo, son adquiridas mediante un proveedor externo. Aunque estas herramientas representan un apoyo en la evaluación crediticia, diversos análisis han demostrado irregularidades en los resultados entregados, afectando la calidad de las decisiones crediticias.

Como se mencionó anteriormente, la dependencia de un proveedor externo limita el control sobre procesos críticos. Esta situación restringe la capacidad de la empresa para realizar mejoras o ajustes rápidos en sus procesos de evaluación crediticia, lo cual puede ser especialmente perjudicial en un mercado financiero que cambia rápidamente.

Además, si las herramientas adquiridas externamente no proporcionan datos precisos, existe el riesgo de aprobar créditos a clientes que no tienen la capacidad real de pago, aumentando así el riesgo de morosidad. El uso de estas herramientas también puede llevar a una mayor exposición a errores sistemáticos que podrían evitarse con una gestión interna de estos procesos.

Por otro lado, el costo constante de estos servicios externos impacta directamente en los gastos operativos de la empresa, reduciendo los recursos disponibles para desarrollar soluciones internas más efectivas y personalizadas. Contar con herramientas propias de gestión crediticia permitiría a la empresa tener un mayor control sobre sus datos, mejorar la precisión de las evaluaciones y reducir la dependencia de terceros, lo que resultaría en una operación más eficiente y ajustada a las necesidades específicas del negocio.

1.3.3. Informe de Deuda Consolidada

Como organismo público encargado de supervisar las actividades financieras, la Comisión para el Mercado Financiero (CMF) implementó un Plan Estratégico durante los años 2020 a 2022, que incluyó cambios normativos destinados a mejorar la información sobre los deudores en el sistema financiero.

Una de las modificaciones más relevantes fue la inclusión de los emisores de tarjetas de crédito no bancarios, como las empresas del retail financiero, en la lista de entidades obligadas a reportar información sobre sus deudores, en cumplimiento de la Ley General de Bancos. Esto permite contar con un historial de pagos de deudas más actualizado y completo, dado el papel crucial que desempeña el retail financiero como herramienta de endeudamiento para los chilenos.

Esta nueva responsabilidad de los emisores de tarjetas no bancarios de informar sobre los deudores y sus respectivas obligaciones financieras también les otorga acceso al informe de deuda consolidada de los individuos en el sistema financiero chileno (código R04). Este informe entrega información de deuda y mora de todas las personas en el sistema financiero chileno, con un periodo de actualización semanal.

El Registro de Deuda Consolidada, administrado por la CMF, incluye información tanto de deudas impagas como de deudas pagadas en el plazo convenido. Este registro tiene como objetivos principales mejorar la transparencia y la calidad de la información crediticia, aumentar la competencia y reducir el sobreendeudamiento. Además, garantiza que el acceso a la información se realice con el consentimiento explícito del deudor y bajo estrictas condiciones de uso y protección de datos.

Esta nueva fuente de datos, sumada a los datos internos recolectados por la compañía tanto en el ámbito financiero como en el retail, sienta las bases para el desarrollo de un modelo interno de Estimación de Ingresos. La disponibilidad de esta información consolidada y actualizada permite a Negocio Financiero realizar evaluaciones de riesgo más precisas y mejorar la gestión del crédito, fortaleciendo así la autonomía y capacidad de respuesta de la empresa en un mercado competitivo.

Capítulo 2: Objetivos

2.1. Objetivo General

El objetivo general del proyecto es desarrollar un modelo de Estimación de Ingresos para el otorgamiento de crédito. Este modelo buscará mejorar la precisión y fiabilidad de las estimaciones de ingresos de los potenciales clientes del Negocio Financiero, permitiendo una gestión de créditos más eficiente y precisa, reduciendo el riesgo de otorgar créditos a personas cuya capacidad de pago real no se corresponda con la información proporcionada.

2.2. Objetivos Específicos

i. Documentar de manera detallada el proceso de extracción y preprocesamiento de datos, culminando en la elaboración de un reporte técnico.

ii. Documentar el algoritmo del modelo de Estimación de Ingresos, explicando su diseño y evaluando su desempeño.

iii. Realizar una validación del modelo con datos históricos y compararlo con las estimaciones de la fuente externa, documentando los resultados en un informe específico.

iv. Proponer estrategias para la integración eficaz del modelo en los procesos de gestión de crédito existentes.

2.3. Alcances

El propósito final del proyecto es generar una estimación de ingresos confiable y precisa, ya que esta información respaldará decisiones críticas, como la aprobación de créditos mediante tarjeta de crédito, buscando mitigar el riesgo al evitar otorgar crédito que exceda la capacidad de pago del individuo. La evaluación del éxito del proyecto no solo se basará en la capacidad del modelo para superar la precisión de las estimaciones actualmente obtenidas de un proveedor externo, sino también en su significancia y representatividad. El modelo debe mostrar un bajo nivel de error y un alto nivel de confiabilidad y precisión.

Las métricas de desempeño, como el error absoluto medio (MAE), se calcularán utilizando el set de ingresos comprobados disponibles para el desarrollo del modelo. Estos indicadores constituirán una medida clave del valor agregado que ofrece la implementación de este modelo interno de estimación de ingresos.

Es importante destacar que el proyecto se centra exclusivamente en el desarrollo del modelo de estimación de ingresos. La fase de desarrollo incluye la recopilación y tratamiento de datos, el entrenamiento y validación del modelo, y la evaluación de su rendimiento en comparación con las estimaciones externas, pero no contempla la implementación operativa del modelo desarrollado.

Capítulo 3: Marco Conceptual

3.1. Machine Learning

El Machine Learning o Aprendizaje automático es una rama de la Inteligencia Artificial que se enfoca en el desarrollo de algoritmos y modelos que permitan a los sistemas aprender patrones y realizar tareas específicas sin necesidad de ser programados explícitamente para cada tarea. En lugar de seguir instrucciones programadas, los modelos de aprendizaje automático utilizan datos para aprender y mejorar con la experiencia.

En el contexto de esta memoria, se utiliza este enfoque debido a su capacidad para manejar conjuntos de datos complejos y extraer patrones significativos de ellos. Este enfoque ha sido ampliamente utilizado en la predicción de ingresos, demostrando ser eficaz en diversas aplicaciones como la estimación y clasificación de niveles de ingresos (Chakrabarty & Biswas, 2018; JPMorgan Chase Institute, 2018).

3.2. Modelos Supervisados

Los modelos supervisados de aprendizaje automático son algoritmos entrenados utilizando conjuntos de datos que incluyen tanto las características de entrada como las etiquetas de salida correspondientes. Estos modelos se emplean para predecir o clasificar nuevas instancias basándose en los patrones aprendidos a partir de los datos de entrenamiento. El principal desafío de esta metodología radica en la calidad y cantidad de los datos de entrenamiento disponibles.

Para el desarrollo del modelo, el entrenamiento utiliza un conjunto de datos que contiene información detallada sobre las características de cada individuo, junto con la etiqueta de salida correspondiente a sus ingresos. Cabe mencionar que el conjunto de datos de ingresos comprobados proviene de las acreditaciones de ingresos de los clientes de Negocio Financiero. Este hecho puede reflejar solo una parte de la población, lo que podría no capturar completamente la diversidad y variabilidad de los ingresos en la población general. Como resultado, el modelo puede aprender patrones sesgados o incompletos y, por lo tanto, puede no generalizar de manera efectiva a nuevas instancias fuera del conjunto de datos de entrenamiento.

3.3. Modelos de Regresión

Los modelos de regresión son una técnica estadística utilizada para predecir valores continuos, como los ingresos de los clientes, en función de variables independientes. En el contexto de la gestión crediticia, donde los ingresos juegan un papel fundamental en la evaluación del riesgo, el uso de este tipo de modelos es preferible a los modelos de clasificación, que buscan predecir categorías.

Los segmentos de interés en la gestión crediticia pueden cambiar con el tiempo debido a modificaciones en las políticas internas de las instituciones financieras o a cambios en el entorno económico. Con un modelo continuo, es posible ajustar las predicciones sin necesidad de redefinir constantemente las categorías de clasificación. Esta flexibilidad permite adaptar las decisiones

crediticias de manera más dinámica, respondiendo mejor a los cambios en el comportamiento del cliente y en las políticas de crédito.

3.1. Gradient Boosting Regressor

Gradient Boosting Regressor es un potente algoritmo de machine learning utilizado para problemas de regresión. Este método se basa en la idea de mejorar un modelo débil (como un árbol de decisión) mediante la combinación de múltiples modelos débiles en una secuencia, de tal manera que cada modelo nuevo intenta corregir los errores cometidos por el modelo anterior.

En Gradient Boosting, cada árbol de decisión se construye secuencialmente, y cada árbol nuevo es entrenado para predecir el residuo del conjunto de árboles anteriores. El objetivo es minimizar una función de pérdida, generalmente el error cuadrático medio (MSE), para mejorar la precisión del modelo.

La fórmula básica del modelo de Gradient Boosting es:

$$F_m(x) = F_{m-1}(x) + h_m(x)$$

donde:

- $F_m(x)$ es el modelo después de la m-ésima iteración.
- $F_{m-1}(x)$ es el modelo después de la m-ésima iteración.
- $h_m(x)$ es el nuevo árbol de decisión que corrige los errores del modelo $F_{m-1}(x)$.

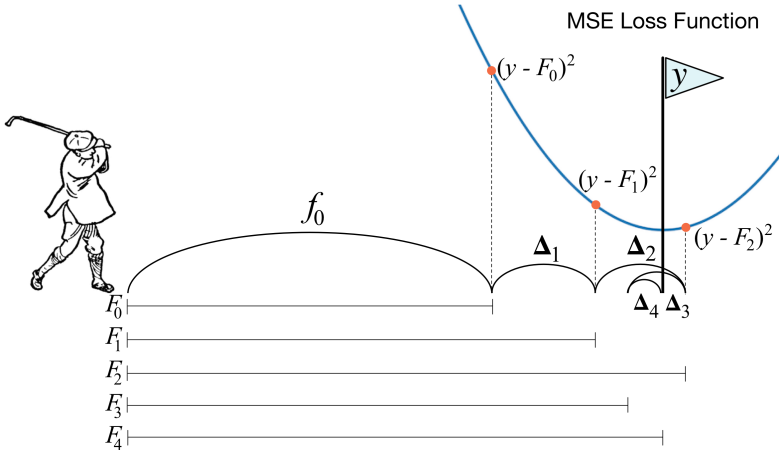
El proceso itera hasta que se alcanza un número predeterminado de árboles o se logra un nivel aceptable de precisión.

La Figura 3 ilustra cómo cada etapa del Gradient Boosting Regressor intenta minimizar el error cuadrático medio (MSE) a lo largo de múltiples iteraciones. La imagen utiliza una analogía con el golf, donde cada swing (iteración) intenta acercarse más al objetivo (minimización del error).

En la imagen, F_0 es el primer modelo, que generalmente es un modelo simple, y cada F_m es un modelo subsiguiente que corrige los errores del modelo anterior. Las diferencias $\Delta_1, \Delta_2, \Delta_3, \Delta_4$ representan la corrección aplicada en cada iteración para acercar las predicciones del modelo al valor verdadero y. El objetivo final es que F_n se acerque lo más posible al valor verdadero de y.

La curva azul representa la función de pérdida MSE (Mean Squared Error), que el modelo busca minimizar. Cada punto en la curva muestra el error cuadrático medio de los diferentes modelos F_m. La idea es que, con cada nueva iteración, el modelo se acerque más al valor verdadero y, reduciendo el error de manera continua.

Figura 3: Ilustración del proceso de Gradient Boosting.



Nota. De How to explain gradient boosting, por T. Parr y J. Howard, 2020, Explained.ai (https://explained.ai/gradient-boosting/).

3.2. Hiperparámetros

Los hiperparámetros son configuraciones establecidas antes de entrenar un modelo de machine learning y tienen un gran impacto en su rendimiento. A diferencia de los parámetros del modelo, que se ajustan automáticamente durante el entrenamiento, los hiperparámetros se definen manualmente.

En este proyecto, los hiperparámetros clave para el modelo Gradient Boosting Regressor fueron:

- Número de estimadores (n_estimators): Indica la cantidad de árboles en el modelo. Un mayor número de estimadores puede mejorar la precisión, pero también aumenta el riesgo de sobreajuste.
- Tasa de aprendizaje (learning_rate): Determina cuánto contribuye cada árbol al modelo final. Una tasa más baja puede aumentar la precisión, pero requiere más árboles.
- Profundidad máxima (max_depth): Define la profundidad de cada árbol. Una mayor profundidad permite capturar relaciones más complejas, pero también puede causar sobreajuste.
- Submuestreo (subsample): Es la fracción de datos utilizada para entrenar cada árbol. Valores más bajos pueden reducir el sobreajuste y mejorar la generalización del modelo
- Muestras mínimas para dividir (min_samples_split): El número mínimo de muestras requeridas para dividir un nodo interno. Un valor más alto puede evitar que el modelo aprenda patrones de ruido en los datos, reduciendo el riesgo de sobreajuste.
- Muestras mínimas en las hojas (min_samples_leaf): El número mínimo de muestras que debe tener un nodo hoja. Este parámetro ayuda a prevenir la creación de nodos hojas con muy pocas muestras, lo que podría llevar a un modelo menos generalizable. Un valor más alto ayuda a suavizar el modelo y a mejorar la capacidad de generalización.

3.3. Validación Cruzada

La validación cruzada es una técnica fundamental en el machine learning utilizada para evaluar y mejorar la capacidad de generalización de un modelo. Su objetivo es asegurar que el modelo no dependa excesivamente de un conjunto específico de datos de entrenamiento y que pueda generalizar bien a datos no vistos.

El proceso de validación cruzada implica dividir el conjunto de datos en k partes iguales, conocidas como folds. En cada iteración, el modelo se entrena utilizando k-1 folds y se evalúa con el fold restante. Este proceso se repite k veces, asegurando que cada fold se utilice como conjunto de prueba una vez. Al final, se promedian las métricas de rendimiento obtenidas en cada iteración para proporcionar una estimación robusta del rendimiento del modelo.

Una de las principales ventajas de la validación cruzada es la reducción del sobreajuste. Al evaluar el modelo en múltiples subconjuntos de datos, es posible detectar y prevenir el ajuste excesivo a un conjunto específico de datos de entrenamiento. Además, permite un uso eficiente de los datos, ya que todo el conjunto de datos se utiliza tanto para entrenamiento como para prueba. Esto es especialmente valioso cuando se dispone de un conjunto de datos limitado.

En este proyecto, se utilizó la validación cruzada k-fold para evaluar y seleccionar los hiperparámetros del modelo Gradient Boosting Regressor. Este enfoque permitió ajustar los hiperparámetros de manera óptima, asegurando que el modelo tenga una buena capacidad de generalización y minimizando el riesgo de sobreajuste. La validación cruzada proporcionó una evaluación confiable del rendimiento del modelo, contribuyendo a su optimización y eficacia.

3.4. Métricas de Rendimiento

En el contexto de la evaluación de las estimaciones otorgadas por el proveedor externo, como también las estimaciones de los modelos desarrollados durante el proyecto, se utilizan distintas métricas estadísticas para evaluar su rendimiento. Doe y Smith (2018) sugieren que el uso de métricas como el MAE y el R^2 son fundamentales para evaluar la precisión de los modelos de estimación de ingresos.

3.4.1. Mean Absolute Error (MAE)

Mide la magnitud promedio de los errores en una serie de predicciones, sin considerar su dirección. Se calcula como la media de las diferencias absolutas entre las predicciones y los valores reales.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

3.4.2. Median Absolute Error (MDAE)

Similar al MAE, pero en lugar de utilizar la media de las diferencias absolutas, utiliza la mediana. Esto hace que la métrica sea menos sensible a valores atípicos.

$$MDAE = \text{median}(|y_1 - \widehat{y_1}|, |y_2 - \widehat{y_2}|, \dots, |y_n - \widehat{y_n}|)$$

3.4.3. Coeficiente de Determinación (R^2)

Indica cuánta variabilidad explica el modelo sobre la variable dependiente. Es un valor entre 0 y 1, donde 1 indica que el modelo explica toda la variabilidad y 0 indica que no explica nada más que la media de los datos.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 , SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

3.5. Transformación Logarítmica

La transformación logarítmica es una técnica utilizada para reducir la asimetría en la distribución de los datos, estabilizar la varianza y manejar valores extremos. Esta técnica es particularmente útil para convertir distribuciones sesgadas positivamente en distribuciones más simétricas, facilitando así el análisis y modelización (Osborne, 2010). Al aplicar esta transformación, los valores extremadamente altos tienen menos impacto en el modelo, lo que mejora la capacidad del algoritmo para capturar patrones subyacentes en los datos.

En este proyecto, se utilizó la transformación logarítmica en las características donde se observó una alta concentración de valores bajos y una cola larga hacia la derecha. Esto permitió obtener distribuciones más simétricas y cercanas a una distribución normal, lo cual es beneficioso para muchos modelos predictivos (Iglesias, Noguera, & Núñez, 2018).

3.6. Z-score

El Z-score es una técnica de estandarización que convierte las características numéricas en una escala común con media cero y desviación estándar uno. Esta técnica es crucial para asegurar que todas las características contribuyan de manera equitativa en modelos sensibles a la escala de los datos, como la regresión logística o las redes neuronales, mejorando su rendimiento (Jain, Duin, & Mao, 2000). La fórmula del Z-score es:

$$z = \frac{x - \mu}{\sigma}$$

Donde x es el valor original, μ es la media de la característica, y σ es la desviación estándar de la característica.

En este proyecto, primero se utilizó la transformación logarítmica para reducir el sesgo y estabilizar la varianza en las variables con distribuciones altamente asimétricas. Posteriormente, se aplicó estandarización mediante el método Z-score para asegurar que las variables transformadas estuvieran en la misma escala, lo que facilita la convergencia de los modelos y mejora la capacidad predictiva en casos donde las diferencias de escala podrían influir negativamente.

El criterio para aplicar primero la transformación logarítmica y luego el Z-score fue garantizar que los modelos pudieran manejar distribuciones sesgadas y variables en diferentes escalas de manera más efectiva, resultando en una mejora global en la precisión predictiva y la estabilidad del modelo.

Capítulo 4: Metodología

4.1. Caracterización de la gestión de créditos

Antes de adentrarse en el desarrollo del Modelo, es esencial realizar una caracterización de la gestión de créditos actual dentro de la empresa. Este proceso implicó un análisis de los procesos existentes, las políticas de evaluación crediticia, los criterios de otorgamiento de crédito y cualquier otra consideración relevante relacionada con la gestión de créditos. Identificar tanto los puntos fuertes como las áreas de mejora en la gestión de créditos actuales es fundamental para el desarrollo del modelo. Este análisis proporciona una base sólida para garantizar que el modelo propuesto se integre efectivamente en los procesos existentes de la empresa.

4.2. Recopilación de Data

La recolección de datos es un paso fundamental en la implementación de cualquier proyecto de modelado predictivo. En este proceso, es crucial seleccionar las fuentes de datos adecuadas que sean escalables y representativas para el desarrollo del modelo.

Para asegurar la viabilidad y la utilidad del modelo, se ha priorizado la inclusión de datasets que cuenten con un número significativo de registros. Esta decisión se basa en la necesidad de disponer de una cantidad suficiente de datos para entrenar y validar el modelo de manera efectiva. Además, se ha considerado fundamental garantizar la representatividad de los datos, de manera que reflejen de manera precisa la diversidad y complejidad de la población objetivo.

4.3. Set de Ingresos y Tratamiento de la Data

En esta fase, se abordará la preparación del set de ingresos mediante el tratamiento adecuado de los datos. Esto incluye la consolidación de registros, la actualización de valores históricos y la transformación de variables para asegurar que los datos estén en condiciones óptimas para el análisis predictivo. Se implementarán técnicas para manejar valores nulos, realizar transformaciones logarítmicas cuando sea necesario y estandarizar las variables, asegurando que el modelo trabaje con datos limpios y precisos.

4.4. Pre Procesamiento de Características

El preprocesamiento de características es un paso crítico para preparar las variables que se utilizarán en el modelo predictivo. Este proceso implica la identificación y tratamiento de características numéricas y categóricas. Se aplicarán técnicas como la imputación de valores nulos, la transformación logarítmica para reducir sesgos en la distribución de datos y la estandarización para normalizar las escalas de las variables. Estas actividades aseguran que las características sean adecuadas y efectivas para el entrenamiento del modelo.

4.5. Análisis de Características

El análisis de características implica evaluar la importancia y relevancia de variables en relación con la estimación de ingresos. Este proceso incluye la exploración de relaciones entre variables, la identificación de patrones y la selección de las características más significativas para el modelo. Se utilizarán técnicas de análisis exploratorio de datos y visualización para entender mejor cómo las diferentes características influyen en los ingresos y para optimizar el modelo predictivo.

4.6. Cálculo de Rendimiento Estimaciones Externas

Antes de iniciar el desarrollo del modelo de estimación de ingresos, es fundamental evaluar el rendimiento de las estimaciones actuales proporcionadas por el proveedor externo. Este proceso implica calcular una serie de métricas de desempeño para comparar las estimaciones externas con el conjunto de ingresos comprobados disponibles. El cálculo de estas métricas proporcionará una base sólida para evaluar la precisión y eficacia del nuevo modelo interno una vez desarrollado, asegurando que cumpla con el objetivo y alcance de mejorar la precisión de las estimaciones de ingresos del proveedor externo.

4.7. Desarrollo del Modelo

En esta etapa central del proyecto, se procede al desarrollo del modelo. Se utiliza como enfoque principal el machine learning, explorando una variedad de algoritmos, desde modelos de regresión hasta redes neuronales, adaptándolos a las necesidades específicas del proyecto. Además, se realizan segmentaciones correspondientes para garantizar una precisión óptima en las estimaciones.

4.8. Test

Esta etapa consiste en evaluar el modelo utilizando un conjunto de datos de prueba independiente, comparando sus predicciones con datos no utilizados durante el entrenamiento para medir su precisión. En este proyecto, se emplearán datos históricos de clientes para esta evaluación. La prueba del modelo es esencial para garantizar su robustez y fiabilidad, proporcionando una medida crucial de su efectividad antes de una posible implementación.

Capítulo 5: Desarrollo y Resultados

5.1. Caracterización de la gestión de créditos

5.1.1. Requisitos Generales Productos Financieros

Los requisitos principales para el otorgamiento de crédito, mediante tarjeta de crédito y créditos de consumo, en Negocio Financiero son los siguientes:

i. Los productos financieros están dirigidos a personas naturales.

ii. Cada producto establece sus propias restricciones de edad mínima y máxima para los solicitantes.

iii. Cada producto tiene su propia base aprobada para el otorgamiento de crédito.

iv. En el caso del crédito de consumo, los solicitantes deben acreditar sus ingresos mediante la presentación de liquidaciones de sueldo u otros documentos pertinentes. Sin embargo, para solicitar una tarjeta de crédito, no se requiere acreditación de ingresos.

5.1.2. Base Aprobada para otorgamiento de Crédito

La base de individuos aprobados para recibir crédito se establece utilizando un Score de Riesgo, conocido internamente como Score de Originación. Este puntaje es proporcionado por el mismo proveedor externo que entrega las estimaciones de ingresos actuales.

En el caso del crédito de consumo, el Score de Originación debe ser igual o superior a 750, mientras que para solicitar una tarjeta de crédito, debe ser igual o superior a 500. La diferencia entre los puntajes, se debe a que el crédito de consumo suele involucrar montos mayores y plazos más largos respecto a la tarjeta de crédito.

5.1.3. Asignación de Cupos y Monto de Crédito

Para la asignación de crédito mediante los productos financieros, se tienen en cuenta factores como el Score de Originación, el ingreso acreditado y los ingresos estimados del solicitante, en el caso de la tarjeta de crédito. Estos criterios ayudan a establecer límites responsables y adecuados para cada individuo, considerando su capacidad de pago y perfil de riesgo.

5.2. Recopilación de Data

5.2.1. Dataset Ingresos Comprobados

El conjunto de datos de ingresos comprobados se extrae de las aplicaciones de créditos de consumo realizadas en Negocio Financiero. Como se detalló en la caracterización de la gestión de créditos, uno de los requisitos de este producto es la acreditación de ingresos mediante liquidaciones de sueldo o documentos pertinentes. Es importante tener en cuenta que esta base de datos tiene el sesgo inherente de las personas que aplican a créditos de consumo en Negocio Financiero. Este sesgo será abordado y estudiado en secciones posteriores para comprender su posible impacto en el desarrollo del modelo.

5.2.2. Dataset Variables Financieras

El informe de deuda consolidada, código R04, se actualiza semanalmente y contiene información detallada sobre las deudas de aproximadamente 11 millones de personas en el sistema financiero. Para enriquecer la estimación de ingresos de los individuos, se ha decidido utilizar un enfoque basado en períodos móviles de 12 meses. Este enfoque permite captar las fluctuaciones en la situación financiera de los individuos a lo largo del tiempo, proporcionando un panorama más representativo de sus niveles de endeudamiento y capacidad de pago.

Dentro de este período móvil de 12 meses, se eligió utilizar el valor máximo de cada variable financiera. Esta elección permite capturar los montos mas altos en las obligaciones financieras de cada individuo, reflejando momentos de máxima carga financiera. Esto resulta útil para modelar la capacidad de pago y, en consecuencia, estimar de manera más precisa los ingresos de los individuos. Las principales variables de este dataset, representadas en miles de pesos, incluyen: créditos directos al día, créditos directos en mora (segmentados por tramo de mora), créditos de consumo, créditos hipotecarios, créditos comerciales, entre otras.

Para determinar el número de lags a utilizar, se realizó un análisis de distintos periodos de retraso (3, 6, 9 y 12 meses) con el objetivo de evaluar cuál capturaba mejor los patrones recurrentes en el comportamiento de la deuda financiera. Los resultados indicaron que el lag de 12 meses era el más adecuado, ya que reflejaba con compeltitud los ciclos de mayor carga financiera, particularmente asociados a eventos estacionales como las festividades y meses de alata carga financiero como lo es marzo. Aunque se probaron lags más cortos, estos no lograron capturar de manera efectiva los patrones estacionales que influyen significativamente en el endeudamiento de los individuos.

5.2.3. Dataset Variables Demográficas

También se ha incorporado información demográfica para enriquecer el análisis y la construcción del modelo de estimación de ingresos. Esta información fue adquirida a través de un proveedor especializado en datos demográficos, garantizando así la precisión y relevancia de los datos utilizados. Además, el universo de estos datos es comparable al del dataset de variables financieras, lo que asegura una alta representatividad y consistencia en el análisis.

Estudios previos han demostrado que factores demográficos pueden influir significativamente en los ingresos de las personas. Por ejemplo, la edad puede estar asociada con el nivel de ingresos y la estabilidad laboral de una persona, proporcionando información sobre la experiencia y la etapa

de la carrera profesional en la que se encuentra. Además, el género puede influir en las oportunidades profesionales y laborales debido a diversas dinámicas sociales (Blau & Kahn, 2017; Guiso, Monte, Sapienza, & Zingales, 2008).

5.2.4. Dataset Variables Retail

La recolección de datos del retail (cadena farmacéutica) representa una oportunidad significativa para enriquecer el desarrollo del proyecto, especialmente considerando la posición de la empresa dentro del sector del retail, lo que constituye una ventaja competitiva importante. La base de datos consiste en un consolidado anual de aproximadamente 8 millones de clientes, proporcionando una visión del comportamiento de compra de los clientes, incluyendo información como el monto transaccional, el número de transacciones realizadas, el grupo socioeconómico al que pertenece cada cliente según las categorizaciones internas del retail, así como la naturaleza de las necesidades que cumplen con sus compras.

La inclusión de esta información es fundamental por varias razones. En primer lugar, ofrece una perspectiva única sobre el comportamiento de consumo de los clientes, permitiendo identificar patrones de gasto, preferencias de compra y comportamientos financieros relevantes para la estimación de ingresos. Además, al ser parte del mismo conglomerado empresarial, se garantiza la coherencia y consistencia de los datos, lo que facilita su integración con otras fuentes de información financiera. Sin embargo, es importante mencionar que esta base de datos fue entregada directamente por el área comercial de la cadena farmacéutica, lo que puede introducir cierta ambigüedad en algunas variables debido a la naturaleza específica del proceso de recopilación y registro de datos.

5.3. Set de Ingresos y Tratamiento de la data

5.3.1. Clientes con más de una Acreditación de Ingresos

Es común que un cliente tenga múltiples acreditaciones de ingresos en distintos años debido a la obtención de nuevos créditos o renegociaciones. Para abordar esto, se eliminaron las acreditaciones duplicadas por cliente, manteniendo solo el registro más reciente para cada individuo. Inicialmente, se identificaron 74.447 acreditaciones duplicadas. Después de eliminarlas, el conjunto de datos se redujo a 92.870 registros únicos.

5.3.2. Actualización de Ingresos

Para abordar el análisis y preprocesamiento del dataset que contiene la variable objetivo, se enfrenta el desafío de que los ingresos acreditados provienen de diferentes años, desde el año 2016 hasta la fecha.

Dado que las emisoras de tarjetas de crédito no bancarias, como en este caso Negocio Financiero, solo disponen de información de deuda consolidada en el sistema financiero a partir de mediados del año 2023, se optó por actualizar los ingresos de años anteriores utilizando dos criterios:

i. Utilizando la última actualización hasta la fecha de variación del ingreso medio anual proporcionada por la Encuesta Suplementaria de Ingresos 2022 (Instituto Nacional de Estadísticas, 2022), se actualizaron los ingresos de años anteriores hasta dicho año.

ii. Posteriormente, se aplicó un factor extra por la variación del Índice de Precios al Consumidor (IPC) acumulado desde 2022 hasta abril de 2024. Para los años 2022 y 2023, el factor de conversión corresponde únicamente al ajuste por IPC, mientras que para 2024 no se aplicó ningún factor de conversión (Instituto Nacional de Estadísticas, n.d.).

Los factores finales de conversión utilizados para la actualización de ingresos de años anteriores se detallan en la Tabla 1.

Tabla 1: Factores de Conversión de Ingresos.

Año	2023	2022	2021	2020	2019	2018	2017	2016
Factor	2,1%	8,7%	12,1%	20,8%	23,8%	26,9%	32,8%	41,7%

Después de realizada la actualización, se determinó trabajar únicamente con los registros a partir del año 2020. Esta decisión se basó en consideraciones sobre la posible inexactitud al actualizar valores muy antiguos. Se estableció un umbral razonable de cuatro años como límite para la inclusión de los datos en el análisis. Finalmente el dataset de ingresos comprobados que se utilizaron para los análisis y desarrollos posteriores cuenta con 45.472 registros.

5.3.3. Métricas y Distribución del Set de Ingresos

Como se mencionó anteriormente, el Set Final de Ingresos está conformado por 45.472 registros. Teniendo en cuenta que la base de datos de deuda consolidada comprende aproximadamente 11 millones de individuos, este conjunto se caracteriza por tener una cantidad relativamente reducida de registros.

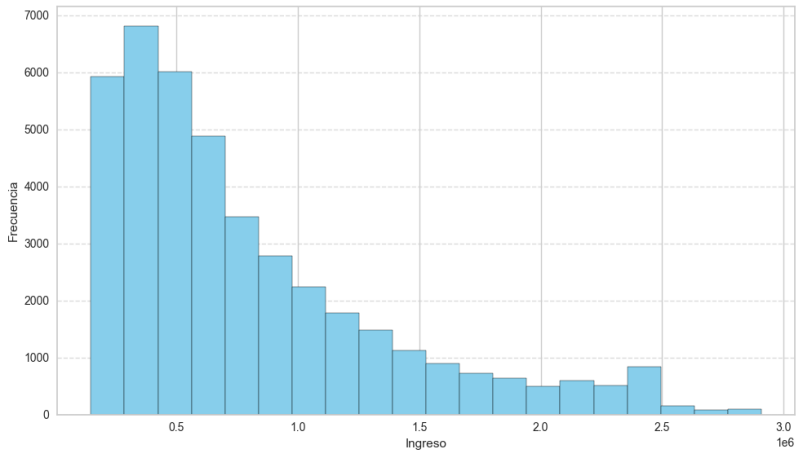
A pesar de esta limitante, se puede observar que la media y la mediana de la muestra (798.986 pesos y 655.000 pesos, respectivamente) son cercanas a las de la población ocupada chilena, que se sitúan en 806.248 pesos y 600.109 pesos según la Encuesta Suplementaria de Ingresos 2022 (Instituto Nacional de Estadísticas, 2022). Esto sugiere que, al menos en términos de tendencia central, la muestra utilizada es parcialemnte representativa de la población general, lo cual permite un mayor grado de confianza en el análisis y aplicabilidad de los resultados posteriores.

Tabla 2: Métricas Set final de Ingresos Comprobados.

Mínimo	142.916
Mediana	655.000
Media	834.238
Máximo	2.914.973

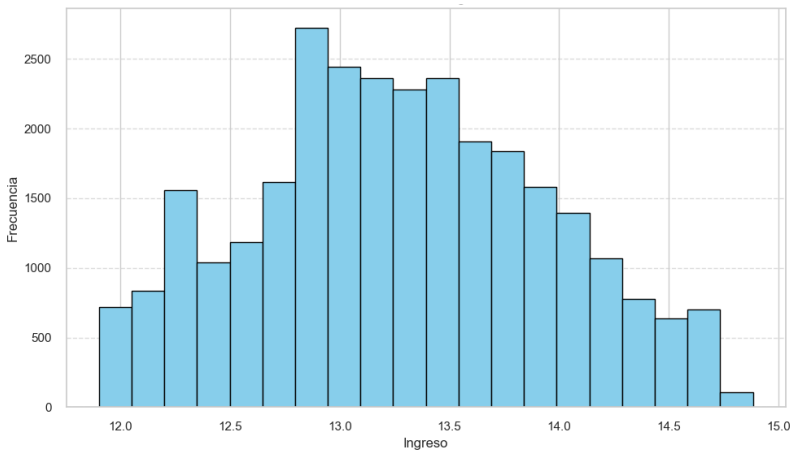
También, como se observa en la Figura 4, la distribución de ingresos muestra una concentración en los rangos bajos, disminuyendo conforme se incrementa el ingreso. Esta tendencia es consistente con la distribución de ingresos observada a nivel nacional, donde una gran parte de la población se encuentra en los segmentos de ingresos más bajos (OECD, 2022; World Bank, 2022). Estos puntos ayudan a establecer que, a pesar de la cantidad relativamente reducida de registros, la muestra utilizada es representativa y adecuada para reflejar la realidad de los ingresos en la población chilena, lo que respalda la validación y aplicación efectiva de modelos y políticas basadas en estos datos.

Figura 4: Distribución Set de Ingreso Final.



Como último ajuste, se aplicó transformación logarítmica a la variable ingreso. Como se observa en la Figura 5, al aplicar la función logaritmo natural, este sesgo se reduce, haciendo que la distribución de los datos sea más simétrica y cercana a una distribución normal.

Figura 5: Distribución Set de Ingresos en escala Logarítimica.



5.4. Pre-procesamiento de Características

5.4.1. Características Numéricas

- **Tratamiento de valores nulos**: Para las características con un alto número de valores nulos, se decidió no utilizarlas en el análisis debido a que su inclusión podría introducir sesgos y errores significativos, comprometiendo la generalizabilidad y precisión del modelo. Esta decisión está respaldada por estudios que indican que la imputación de datos faltantes en exceso puede deteriorar la calidad del modelo predictivo (Rubin, 2004; Little & Rubin, 2019).

- **Transformación Logarítmica:** Las características se transformaron utilizando la función logarítmica para mejorar su rendimiento en el desarrollo de los modelos de machine learning. Específicamente, se aplicó la transformación log(x+1) para poder tratar valores iguales a cero sin generar valores indefinidos.

- **Estandarización:** Las características numéricas fueron estandarizadas mediante el método z-score, con el objetivo de garantizar que los algoritmos desaroollados no se vieran afectados por las diferencias en la escala de las variables

5.4.2. Características Categóricas

- **Codificación Binaria:** Se transformaron las características en variables binarias para facilitar su análisis en modelos que requieren entradas numéricas. Por ejemplo, la variable de género fue codificada como 0 para hombres y 1 para mujeres, simplificando su incorporación en el análisis y desarrollo del modelo.

- **Re-Categorización:** Se estudiaron las categorías de las variables y se decidió recategorizar cuando algunas tenían comportamientos similares o no eran suficientemente representativas. Esto permitió consolidar la información y mejorar la calidad del análisis.

- **Creación de Variables:** Se generaron nuevas variables derivadas de las existentes para capturar mejor la información relevante. Por ejemplo, se creó una variable que indica si una persona está jubilada, utilizando las variables de edad y género. Estas nuevas variables ayudan a enriquecer el análisis al considerar factores adicionales importantes.

5.5. Análisis de Características

5.5.1. Características Financieras

Como se especificó anteriormente, las características financieras provienen del archivo código R04, proporcionado por la Comisión para el Mercado Financiero (CMF), que contiene información detallada sobre el endeudamiento de los individuos en el sistema financiero. A continuación, se presentará el análisis de las características financiera utilizando como ejemplo la variable de monto adeudado en créditos de consumo (CRE_CONSUMOS).

5.5.1.1. Distribución

Como se observa en la Figura 6, la distribución de la variable presenta un fuerte sesgo positivo. Esto indica que la mayoría del monto adeudado por las personas se sitúa en valores relativamente bajos, hasta los 10 millones de pesos, representando aproximadamente el 75% del área bajo la curva de densidad. Esta característica es crucial para entender el comportamiento del endeudamiento en el sistema financiero.

El fuerte sesgo positivo sugiere que, aunque existen individuos con deudas significativamente altas, la mayoría de los clientes del set de ingresos tienen deudas de consumo que no superan los 10 millones de pesos. Además, podemos como ilustra la Figura 7, la distribución de la deuda de consumo para la población total del sistema financiero, también acumula un 75% de la muestra en torno de los 10 millones de pesos. Esto es fundamental para garantizar la representatividad de la muestra respecto a la población total. Este patrón es común en el análisis de datos financieros, donde una minoría de clientes tiende a tener deudas excepcionalmente altas, mientras que la mayoría se mantiene dentro de un rango más manejable.

Este comportamiento también se observa en las demás variables financieras analizadas, como créditos directos al día, créditos directos en mora y créditos hipotecarios, las cuales muestran una concentración de valores en la parte izquierda de la distribución, con una larga cola hacia la derecha.

Figura 6: Distribución variable CRE_CONSUMOS en Set de Ingresos Comprobados.

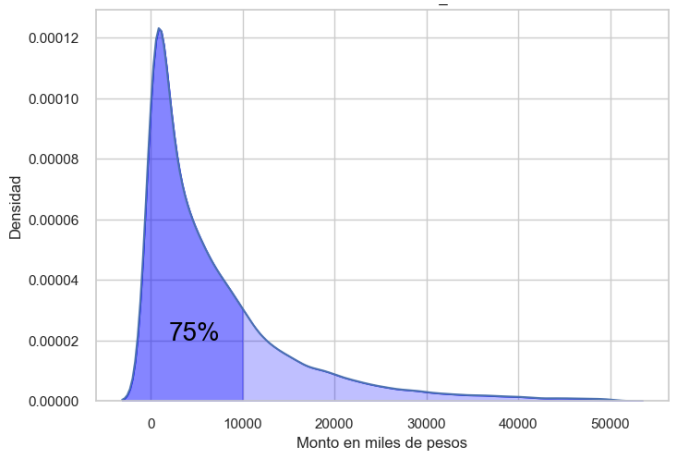
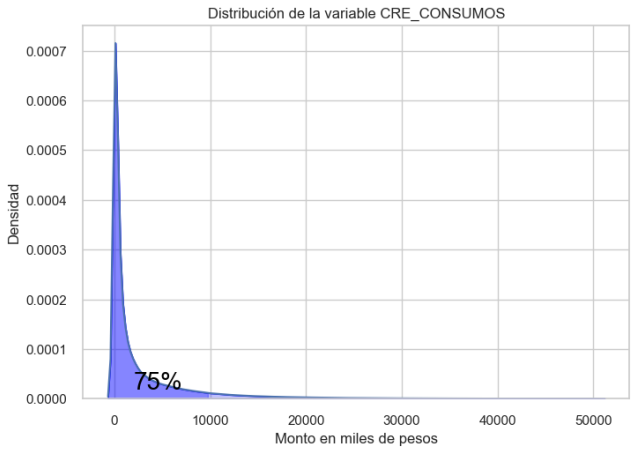


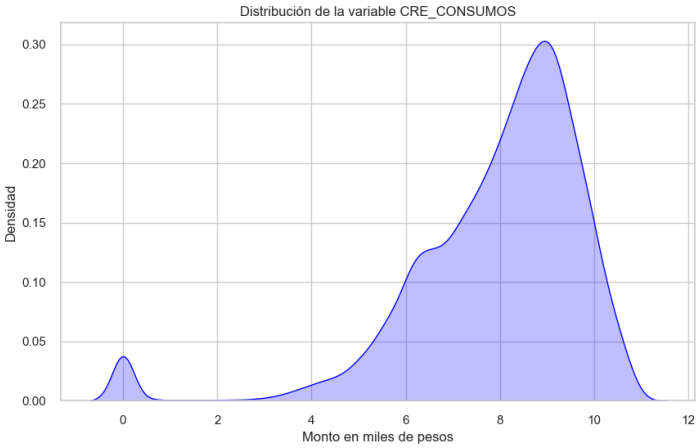
Figura 7: Distribución variable CRE_CONSUMOS población total en Sistema financiero.



Para mejorar la calidad de los datos y la precisión del modelo predictivo, se aplicó transformación logarítmica a las variables financieras, mismo ajuste utilizado con el conjunto de ingresos comprobados. En particular, para la variable CRE_CONSUMOS, la transformación logarítmica tuvo un impacto significativo en la reducción del sesgo. El valor del sesgo se redujo en más de un 50% en valor absoluto, mejorando notablemente la simetría de la distribución.

La Figura 8 muestra la distribución de la variable después de la transformación logarítmica. Como se puede observar, la transformación ha logrado concentrar los datos y reducir la cola larga hacia la derecha, proporcionando una distribución más manejable y adecuada para el análisis y modelado predictivo.

Figura 8: Distribución variable CRE_CONSUMOS.



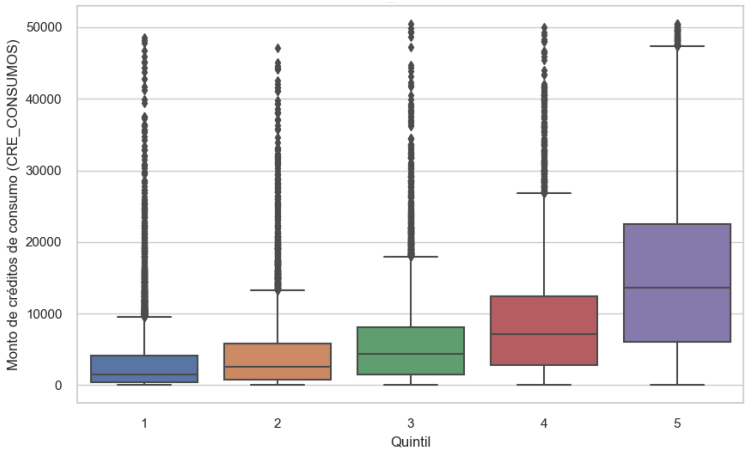
5.5.1.2. Análisis por quintil de ingresos

El análisis por quintil de ingresos de la muestra, representado en la Figura 9, proporciona una visión general de cómo se distribuye el monto adeudado en créditos de consumo entre diferentes segmentos de ingresos. Cada Boxplot representa un quintil, lo que permite comparar la dispersión y centralización del monto de deuda entre los distintos segmentos de la población.

Observamos que los quintiles inferiores (1 y 2) tienen una mediana de deuda significativamente menor en comparación con los quintiles superiores. Esto indica que los individuos en estos quintiles tienden a tener deudas de consumo menores, lo cual puede estar correlacionado con ingresos más bajos.

A medida que avanzamos hacia los quintiles superiores, especialmente el quinto quintil, notamos un aumento notable en la mediana y una mayor dispersión de los datos. La dispersión en los quintiles superiores (4 y 5) es considerablemente mayor, lo que sugiere una variabilidad significativa en los montos de deuda dentro de estos grupos. Esto refleja que los individuos en los quintiles más altos no solo tienen acceso a mayores créditos de consumo, sino que también exhiben una mayor diversidad en los montos adeudados.

Figura 9: Boxplot variable CRE_CONSUMOS por Quintil de ingresos



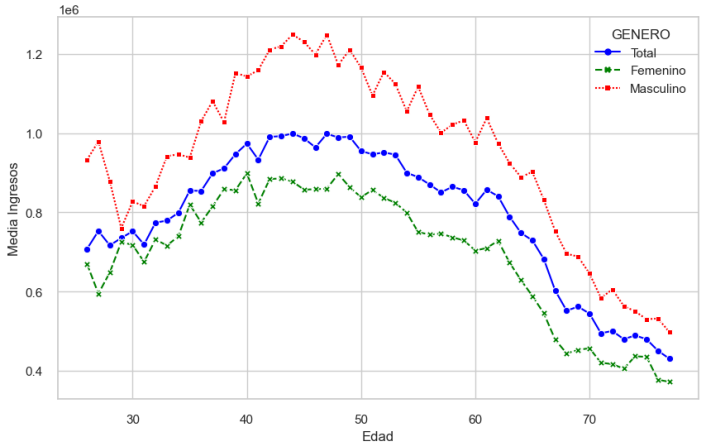
5.5.2. Características Demográficas

5.5.2.1. Edad

La variable de edad, construida a partir de la fecha de nacimiento del individuo, muestra un comportamiento esperado que refleja las tendencias generales observadas en la población. Como se observa en la Figura 10, hay un crecimiento constante en los ingresos hasta aproximadamente los 40 años de edad. Posteriormente, tiende a mantenerse más o menos estable hasta alrededor de los 50 años, momento en el cual comienza a disminuir a un ritmo constante hasta los 60 años, donde aumenta notablemente la tasa de disminución de ingresos.

Se concluye que la muestra presenta un comportamiento similar al de la población general, reflejando las tendencias de los ingresos por edad, como la jubilación y la reducción de ingresos que ocurren en esta etapa de la vida (OECD, 2023).

Figura 10:Tendencia de Ingreso por Edad

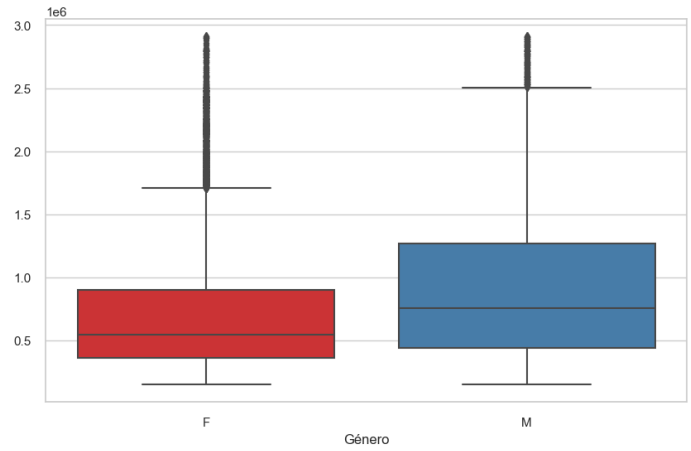


5.5.2.2. Género

El análisis por género revela diferencias significativas en varios aspectos clave. En primer lugar, se observa una marcada disparidad en la proporción de género en la muestra de datos, con un 62% de registros correspondientes a individuos de género femenino y un 38% a individuos de género masculino.

En cuanto a las medidas de tendencia central, se aprecian notables discrepancias entre los ingresos por género, como se observa en la Figura 11. Tanto la media como la mediana de renta para los individuos de género masculino es superior al género femenino, en un 25% y 28% respectivamente. Esta brecha es consistente con la información disponible por la Encuesta Suplementaria de Ingresos, que muestra disparidades salariales significativas entre hombres y mujeres en Chile (INE, 2023).

Figura 11: Boxplot variable Género

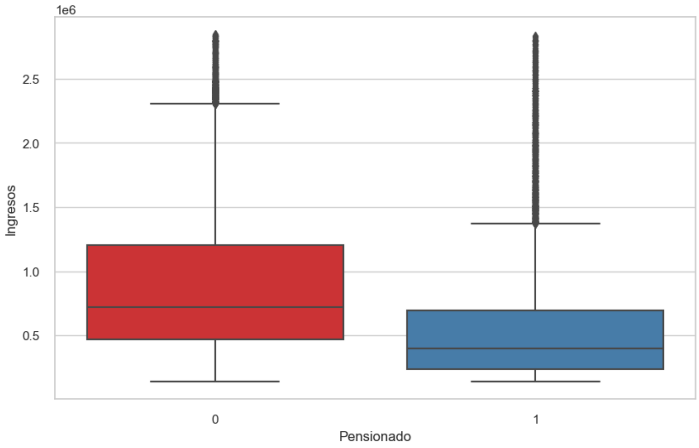


5.5.2.3. Pensionado

Según información proporcionado por el INE en abril del 2024, la edad media de jubilación en Chile para el género femenino es de 62 años, mientras que para el género masculino es de 66 años (Unholster & La Tercera, 2024). Haciendo uso de esta información, y de las variables Edad y Género que se disponen, se crea la variable pensionado.

Como se observa en el gráfico de caja de la Figura 12, la creación de esta variable logra capturar diferencias significativas en la media de renta para cada clase, en donde los pensionados tienen una media 40% inferior que los individuos no pensionados. Como se mencionó anteriormente, esto tiene lógica debido a que los ingresos de jubilados son significativamente menores al resto de la población adulta.

Figura 12: Boxplot variable Pensionado



5.5.3. Características Retail

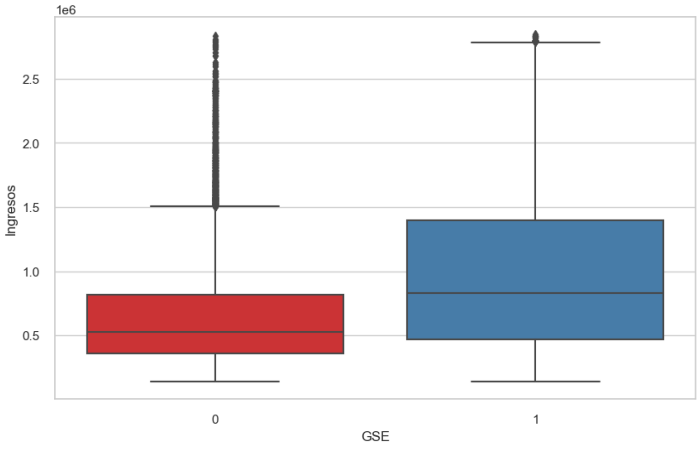
En la siguiente sección se presenta el análisis y pre procesamiento de las principales variables obtenidas de las transacciones de los clientes en la cadena farmacéutica. Como se mencionó anteriormente, la data consiste en un consolidado anual de las transacciones en tienda física de la farmacia, con un universo aproximado de 8 millones de individuos.

5.5.3.1. GSE Retail

El sector retail de la compañía asigna una clasificación socioeconómica a cada cliente utilizando criterios específicos internos. Los grupos socioeconómicos son: E, D, C3, C2, ABC1 y sin clasificación. La mayoría de los individuos pertenece a los segmentos socioeconómicos medios y bajos, específicamente los grupos C3, D y E, representando el 87% del total de la muestra estudiada. Esta distribución es coherente con la importancia del retail financiero en el acceso a crédito en los segmentos socioeconómicos más bajos del país (Valenzuela Aros, 2016; Palacios, 2020).

Al analizar cómo varían los ingresos entre los distintos segmentos, se observa que, aunque la clasificación socioeconómica del retail no ordena perfectamente el promedio de ingresos de los individuos, sí permite distinguir dos segmentos principales: los tres grupos más altos (AB1, C2 y C3) y los grupos más bajos (D, E y sin clasificación). En consecuencia, se optó por crear dos clases que representan estos segmentos. Como se puede apreciar en el gráfico de caja de la Figura 13, existe un contraste significativo en el comportamiento de ingresos entre estas dos clases, destacando diferencias importantes en los niveles de ingreso entre ellos.

Figura 13: Gráfico de caja variable GSE



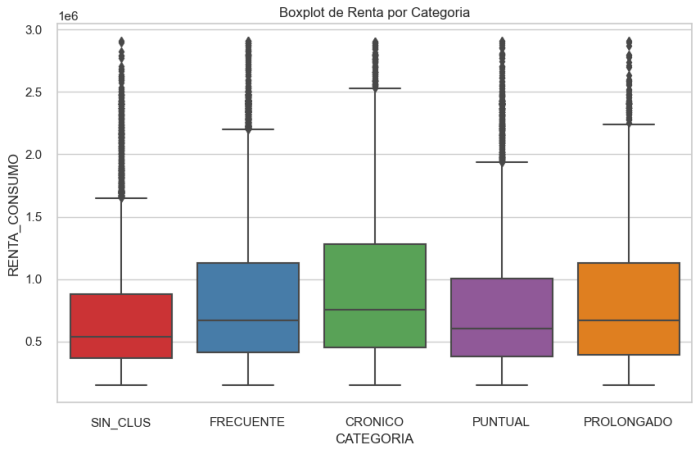
5.5.3.2. Necesidades

Esta variable especifica la necesidad que cumple el cliente con sus compras en la farmacia, en donde prevalece la necesidad crónica, que se refiere a cuando un cliente compra medicamentos asociados a enfermedades crónicas en la farmacia. Como se observa en el gráfico de cajas de la Figura 14, los clientes catalogados con necesidad Crónico tienen un promedio de ingreso mayor que el resto.

Es importante señalar que la cadena farmacéutica en cuestión no tiene convenios con Fonasa ni con la Central Nacional de Abastecimiento (CENABAST), lo que implica que sus clientes no pueden beneficiarse de los descuentos y facilidades que estos sistemas públicos de salud ofrecen para la compra de medicamentos. Fonasa ofrece descuentos significativos en medicamentos para enfermedades crónicas a través de convenios con diversas farmacias, en el marco de políticas como la Ley CENABAST (Ministerio de Salud, 2020). Como resultado, el perfil de los clientes que compran medicamentos crónicos en esta cadena farmacéutica tiende a ser de ingresos más altos. Esto se debe a que estos clientes están dispuestos y son capaces de pagar precios más elevados por los medicamentos que necesitan.

Además, se observa que las necesidades Frecuente y Prolongado tienen un comportamiento similar entre ellas, y lo mismo con las categorías Puntual y Sin Clúster. De esta manera se crearon tres clases: Crónico, Frecuente/Prolongado y Puntual/Sin Clúster. Esta segmentación logra capturar de manera más clara y concisa, reduciendo la dimensionalidad de los datos.

Figura 14: Boxplot variable Necesidades



5.6. Cálculo Rendimiento Estimaciones Externas

Se analiza la precisión de las estimaciones de ingresos proporcionadas por el proveedor externo utilizando el conjunto de datos de ingresos comprobados. El objetivo principal es evaluar la exactitud de estas estimaciones para los individuos en dicho conjunto e identificar oportunidades de mejora en el proceso de estimación. Las métricas calculadas incluyen el error medio absoluto (MAE), el error mediano absoluto (MDAE), y el coeficiente de determinación (R^2), presentadas en la Tabla 3 y Tabla 4.

Tabla 3: Métricas de Rendimiento de Estimaciones de Ingresos Externas para el Año 2024.

MAE	416.825
MDAE	270.642
R2	0.08

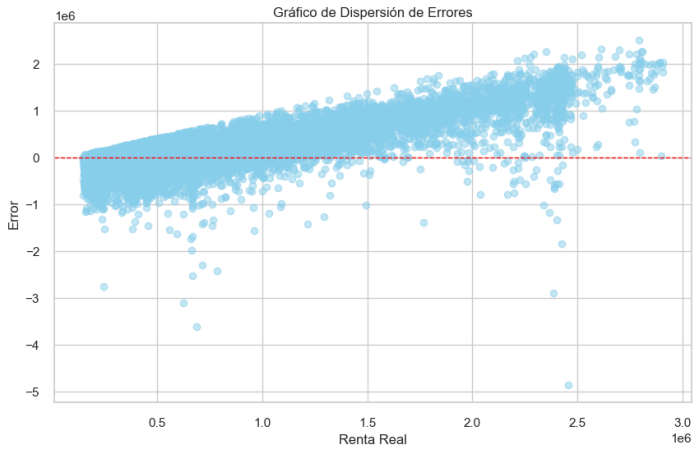
Tabla 4: Métricas de Rendimiento Estimaciones de Ingresos Externas para el Año 2020 a 2023.

MAE	357.715
MDAE	226.755
R2	0.19

Considerando el promedio del set de ingresos comprobados, podemos observar que el MAE es bastante alto, llegando a estar sobre los 400 mil pesos para los registros provenientes del año 2024. Estos resultados presentan un gran espacio de mejora para el desarrollo del modelo de estimación de ingresos.

Como se observa en la Figura 15, el error tiende a aumentar a medida que los ingresos se incrementan. Específicamente, todos los puntos por encima del eje horizontal representan subestimaciones, lo que indica que las estimaciones tienden a subestimar los ingresos superiores al millón de pesos.

Figura 15: Gráfico de dispersión de errores para el Set de Ingresos del año 2024



Este análisis explica la diferencia en los errores observados en la Tabla 4 y Tabla 5. El promedio de ingresos comprobados para la muestra del año 2024 es de 834.334 pesos, mientras que para los ingresos correspondientes a los años 2020 a 2023 es de 723.678 pesos. Esta diferencia en los promedios justifica el menor rendimiento (mayor error absoluto) de las estimaciones externas para los datos más recientes.

5.7. Desarrollo del Modelo

5.7.1. Data Usage

Como primer paso para el desarrollo del modelo, fue necesario definir el uso de los datos para las distintas etapas del proyecto: entrenamiento, validación y testeo. Disponiendo de datos desde el año 2020 hasta el año 2024, se decidió separar los datos en dos conjuntos principales: los registros del período 2020-2023 y los registros del año 2024.

La primera base de datos, que comprende los registros desde 2020 hasta 2023, se utilizó para el entrenamiento y validación del modelo. Durante la fase de entrenamiento, se implementó la técnica de validación cruzada. Esta técnica permitió dividir los datos de entrenamiento en múltiples subconjuntos, usando algunos para entrenar el modelo y otros para validarlo, de manera iterativa. Esto optimizó los hiperparámetros del modelo y ayudó a prevenir el sobreajuste al evaluar su rendimiento en diferentes particiones de los datos.

Para el testeo final, se utilizó exclusivamente la data del año 2024, manteniéndola separada durante todo el proceso de desarrollo del modelo. Este conjunto de datos se utilizó para evaluar el rendimiento del modelo en un contexto completamente nuevo, proporcionando una medida más realista de su capacidad para generalizar y desempeñarse en escenarios futuros no anticipados. Finalmente, la cantidad de registros para cada conjunto de datos se presenta en la Tabla 5.

Tabla 5: Cantidad de Registros para Sets

Set	Cantidad de Registros
Entrenamiento / Validación	39.908
Testeo	5.564

5.7.2. Segmentación

Para la segmentación del modelo, se optó por utilizar la cantidad de datos disponibles por individuo como criterio principal. Como se mencionó previamente, la cantidad de registros de clientes de la farmacia es menor en comparación con el total de personas en el sistema financiero, lo cual constituye una limitación significativa. En respuesta a esta limitación, se decidió entrenar dos modelos diferenciados:

i. Modelo Financiero: Este modelo utiliza información demográfica y el informe de deuda consolidado. La ventaja de este enfoque es que puede abarcar un universo amplio de datos, permitiendo la inclusión de aproximadamente 11 millones de individuos.

ii. Modelo Farmacia: Además de la información demográfica y de deuda consolidada, este modelo incorpora información transaccional específica de los clientes de la farmacia. Aunque este enfoque reduce el tamaño del universo a aproximadamente 8 millones de individuos, añade una capa adicional de detalle que puede mejorar la precisión de las estimaciones de ingresos para este segmento específico.

5.7.3. Entrenamiento

5.7.3.1. Selección Algoritmo

Durante el entrenamiento de los modelos, se evaluaron más de 50 algoritmos de regresión diferentes de Machine Learning. Para seleccionar el mejor algoritmo, se utilizó el MAE como métrica principal de rendimiento, con el objetivo de minimizar este valor y así mejorar la precisión predictiva del modelo. Para ambos modelos segmentados, los mejores tres algoritmos de ML fueron: Gradient Boosting Regressor (gbr), Light Gradient Boosting Machine (lightgbm) y Extreme Gradient Boosting (xgboost).

Estos tres algoritmos son del tipo Boosting, es decir, funcionan combinando múltiples modelos débiles, típicamente árboles de decisión, para formar un modelo fuerte y mejorar la precisión de las predicciones. En el caso del proyecto, este tipo de algoritmos tiene el mejor rendimiento debido a varias razones. Primero, su capacidad para reducir el sesgo y la varianza a través de la combinación iterativa de modelos débiles mejora significativamente la precisión. Segundo, los algoritmos de Boosting son eficaces en el manejo de datos con distribuciones sesgadas y pueden captar sutiles interacciones entre variables. Tercero, estos algoritmos implementan técnicas avanzadas de regularización y optimización que ayudan a prevenir el sobreajuste y mejoran la generalización en datos de prueba (Chen & Guestrin, 2016; Friedman, 2001; Ke et al., 2017).

5.7.3.2. Ajuste Hiperparámetros

Debido a tener el mejor rendimiento, se continuó el desarrollo utilizando Gradient Boosting Regressor para ambos modelos. Para encontrar los hiperparámetros óptimos, se utilizó la optimización bayesiana. Inicialmente, se definió un amplio espacio de búsqueda para explorar diferentes configuraciones de hiperparámetros. A través de múltiples iteraciones de validación cruzada, el modelo fue ajustado progresivamente, reduciendo el espacio de búsqueda en cada iteración. Este enfoque permitió identificar las configuraciones que ofrecían el mejor rendimiento generalizado. Después de varias iteraciones, se alcanzó una configuración óptima de hiperparámetros para cada modelo, detallados en la Tabla 6, que maximizó la precisión del modelo mientras se mantenía su capacidad de generalización.

Tabla 6: Valores óptimos Hiperparámetros gbm.

Hiperparámetros	Modelo Financiero	Modelo Farmacia
learning_rate	0.0229	0.056
max_depth	5.0	3.0
n_estimators	348.0	228.0
min_samples_split	9.0	7.0
min_samples_leaf	4.0	4.0
subsample	0.83	0.55

5.7.3.3. Resultados Entrenamiento

En la Tabla 7 se presentan los resultados del entrenamiento para ambos modelos. Los valores han sido transformados a la escala original de los ingresos. Se puede observar que ambos modelos tienen un mejor rendimiento que la estimación externa, lo que demuestra la eficacia de los modelos desarrollados en este proyecto. Además, se observan mejoras marginales al incluir variables provenientes de las transacciones en la farmacia, lo que sugiere que estas variables adicionales contribuyen positivamente a la precisión del modelo.

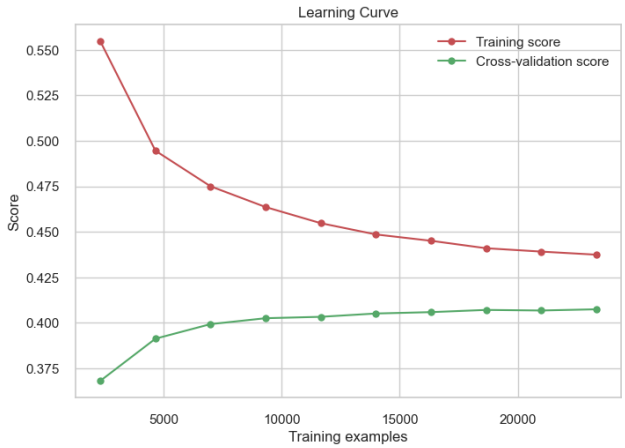
Tabla 7: Resultados Modelos en Entrenamiento.

Estimación	MAE	MDAE	R2
Externa	357.715	226.755	0.19
Modelo Financiero	276.755	162.323	0.38
Modelo Farmacia	269.540	153.786	0.41

Para visualizar mejor el rendimiento de los modelos, se incluye la curva de aprendizaje del Modelo Financiero en la Figura 16. Inicialmente, la puntuación en el conjunto de entrenamiento es alta, indicando que el modelo se ajusta bien a los datos de entrenamiento. Sin embargo, a medida que aumenta el número de ejemplos de entrenamiento, la puntuación disminuye ligeramente y se estabiliza, sugiriendo que el modelo está generalizando mejor y evitando el sobreajuste.

Por otro lado, la puntuación en el conjunto de validación cruzada mejora constantemente a medida que se incrementan los ejemplos de entrenamiento, alcanzando un punto de estabilización. Esto indica que el modelo está aprendiendo y generalizando adecuadamente, beneficiándose de la mayor cantidad de datos. Esta mejora continua en la puntuación de validación cruzada refuerza la robustez del modelo y su capacidad de generalización.

Figura 16: Curva de Aprendizaje Modelo Financiero.



5.8. Test

Los resultados del test muestran que ambos modelos desarrollados son robustos y superan significativamente el rendimiento de las estimaciones proporcionadas por el proveedor externo. Esto se puede observar en la Tabla 8, donde se comparan las métricas de error absoluto medio (MAE), error absoluto mediano (MDAE) y el coeficiente de determinación (R^2).

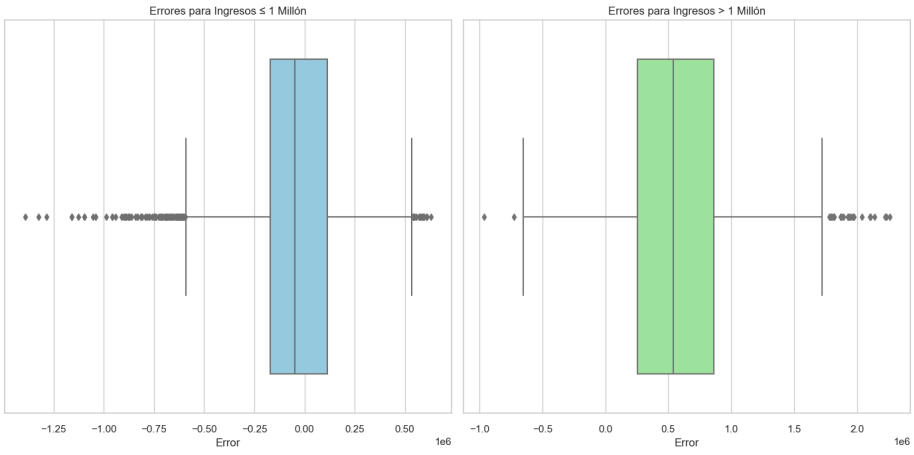
Tabla 8: Resultados Modelos en Test.

Estimación	MAE	MDAE	R2
Externa	416.825	270.642	0.08
Modelo Financiero	318.127	207.901	0.34
Modelo Farmacia	316.338	199.240	0.37

Ambos modelos muestran una reducción significativa en el MAE y MDAE, además de un aumento en el R^2 en comparación con las estimaciones externas. El Modelo Farmacia, que incluye variables transaccionales de los clientes de la farmacia, presenta un rendimiento marginalmente superior al Modelo Financiero.

No obstante, los errores en los ingresos del año 2024 son mayores que los de los datos de 2020 a 2023. Este fenómeno indica la persistente dificultad en la predicción de ingresos más altos, lo cual es consistente con la naturaleza volátil y variada de los ingresos en este segmento. Como se observa en la Figura 17, cuando se desglosan los errores por ingresos superiores e inferiores al umbral del millón de pesos, el error en el segmento superior es significativamente mayor, elevando el error promedio. Sin embargo, dado que el principal público objetivo del Negocio Financiero se centra en clientes con ingresos inferiores al millón de pesos, el modelo está adecuadamente ajustado para satisfacer las necesidades de la compañía.

Figura 17: Boxplots de errores para Ingresos sobre y bajo 1 millón de pesos Modelo Financiero.



Capítulo 6: Discusiones

6.1. Limitaciones del Modelo y los Datos

Uno de los principales desafíos del proyecto es la representatividad de los datos. Aunque se ha logrado una reducción considerable del error en la estimación de ingresos comparado con el proveedor externo, el conjunto de datos utilizado presenta un sesgo inherente al estar compuesto principalmente por clientes que han solicitado crédito en Negocio Financiero. Este sesgo podría limitar la capacidad del modelo para generalizar a clientes potenciales que no forman parte de esta base, lo que podría llevar a subestimaciones o sobreestimaciones del riesgo.

Otra limitación importante radica en la actualización de los ingresos pasados utilizando factores como el IPC. Aunque este ajuste proporciona cierta corrección por inflación, no captura la complejidad de los ingresos individuales en tiempos de alta volatilidad económica. En consecuencia, el modelo podría estar siendo entrenado con ingresos no del todo correctos, lo que afectaría su capacidad predictiva.

De esta manera, la ampliación del set de ingresos, particularmente con datos más recientes, es crucial para mejorar la capacidad predictiva del modelo. Ampliar el set de ingresos comprobados, mediante fuentes de datos tanto internas como externas, permitiría capturar una mayor diversidad en los perfiles económicos, mejorando la precisión y generalización del modelo.

6.2. Desarrollo de Herramientas Internas y Expansión hacia la Gestión Integral del Riesgo Crediticio

El desarrollo del modelo de estimación de ingresos representa un hito clave para la creación de herramientas internas de gestión crediticia dentro de la compañía. El éxito del proyecto ha demostrado que es posible desarrollar soluciones propias utilizando los recursos disponibles. Esta independencia permite a la empresa adaptarse rápidamente a cambios del mercado o nuevas políticas, facilitando la innovación continua en la toma de decisiones.

Como se mencionó anteriormente, el score de riesgo para la originación de clientes se adquiere mediante un proveedor externo, lo que limita la capacidad de personalización de los criterios de evaluación. Este proyecto sienta las bases para el desarrollo interno de esta herramienta, permitiendo a la empresa adaptar el score de acuerdo con sus propias políticas y necesidades, mejorando la precisión y reduciendo los costos asociados a largo plazo.

Otra extensión interesante de este proyecto sería explorar la posibilidad de determinar directamente la probabilidad de default (PD) utilizando las características estudiadas en el modelo de estimación de ingresos. La información obtenida a partir de las variables demográficas, financieras y de ingresos podría ser aprovechada para modelar con mayor precisión la probabilidad de que un cliente caiga en incumplimiento. Al construir un modelo que incorpore estas variables, sería posible especular sobre el riesgo de default de los clientes en distintos horizontes temporales, ajustando las predicciones en función de condiciones macroeconómicas o cambios en el comportamiento del cliente.

Además, una extensión natural de esta predicción sería la estimación de las pérdidas asociadas (Loss Given Default, LGD). La evaluación de las pérdidas en caso de incumplimiento permitiría a la institución no solo anticipar el riesgo, sino también cuantificar el impacto financiero de esos defaults, lo que sería crucial para mejorar la gestión de reservas y provisiones.

6.3. Propuestas Estratégicas a Futuro

Desde una perspectiva estratégica, la Ley REDEC desempeña un papel crucial en el futuro de la gestión crediticia. Esta legislación fomenta un acceso más amplio y transparente a los datos financieros, al ampliar las entidades financieras que deben reportar a la CMF. Esto permitirá a las instituciones financieras desarrollar modelos más precisos y representativos, al incorporar una mayor diversidad de fuentes de información financiera. La implementación de herramientas internas debe tener en cuenta estos cambios regulatorios, asegurando que la empresa pueda adaptarse rápidamente para cumplir con las nuevas exigencias del mercado y aprovechar las oportunidades que estas normativas presentan.

Asimismo, la Ley de Finanzas Abiertas, como parte de la reciente Ley Fintech, establece un marco regulatorio que facilita el intercambio eficiente de información financiera entre proveedores de servicios financieros (CMF Chile, 2024). Este entorno regulatorio crea oportunidades significativas para Negocio Financiero, permitiendo abordar la asimetría de información y diseñar productos más personalizados. La integración de estos principios fortalecerá la capacidad de la empresa para innovar continuamente y adaptarse a nuevas normativas, asegurando que esté bien posicionada para aprovechar las oportunidades que ofrece el entorno en constante evolución.

Por otro lado, la recolección de más variables retail provenientes de empresas del holding, específicamente de la cadena farmacéutica, ofrece una oportunidad invaluable para enriquecer el modelo. La inclusión de variables retail ha mostrado mejoras marginales, lográndose identificar un gran potencial en la obtención de más datos comerciales. Sin embargo, actualmente el acceso a esta información está limitado por la falta de comunicación directa con el área retail de la cadena farmacéutica. Establecer una colaboración más estrecha entre unidades del holding permitiría mejorar significativamente la estimación de ingresos, aumentando la precisión del modelo y facilitando el desarrollo de nuevas herramientas, como campañas de marketing más efectivas y productos financieros más personalizados.

Finalmente, la recalibración continua del modelo es imprescindible. A medida que cambian las condiciones económicas y el comportamiento de los clientes, el modelo debe ajustarse automáticamente para mantenerse efectivo. La creación de pipelines automáticos para la recalibración y la incorporación de retroalimentación constante asegurarán que el modelo siga siendo relevante y preciso a lo largo del tiempo. Este enfoque adaptativo permitirá a la institución mantenerse a la vanguardia en la gestión del riesgo crediticio, ajustando sus estrategias en función de los cambios en el mercado y en los comportamientos de los clientes.

Capítulo 7: Conclusiones

El desarrollo del modelo de estimación de ingresos representa un avance significativo en la gestión del riesgo crediticio dentro de Negocio Financiero. Al reducir en más de un 25% el Error Absoluto Medio (MAE) respecto a las estimaciones externas, el modelo ofrece una mayor precisión y fiabilidad para las decisiones crediticias, disminuyendo el riesgo de default y morosidad. Esta mejora no solo se traduce en una mejor asignación de cupos crediticios, sino que también sienta una base sólida para el desarrollo de herramientas internas personalizadas para la evaluación de riesgo.

La creación de herramientas internas permite a la empresa adaptarse rápidamente a cambios del mercado y ajustar sus políticas crediticias de manera eficiente. Esto también fomenta la innovación continua dentro de la organización y refuerza su competitividad en un entorno financiero cada vez más dinámico y competitivo. Además, como se destacó en la sección de discusiones, el score de originación actualmente adquirido a un proveedor externo también puede ser reemplazado o complementado con soluciones internas. Este aspecto es clave para asegurar la independencia de la empresa en la toma de decisiones estratégicas a largo plazo y personalizar el score de acuerdo con las necesidades particulares de su negocio.

Por otra parte, el modelo plantea nuevas oportunidades para mejorar las metodologías tradicionales de análisis crediticio, incluyendo la posibilidad de determinar directamente la probabilidad de default (PD) de acuerdo con las características estudiadas, así como la pérdida asociada a esos defaults (LGD). Esta capacidad permitiría a la empresa optimizar sus estrategias de mitigación de riesgos y ajustar los criterios de asignación de crédito de manera más precisa y proactiva.

Finalmente, el éxito de este modelo establece un precedente para futuras aplicaciones, como la incorporación de variables adicionales provenientes del retail y otros sectores del holding, lo que podría ampliar aún más su capacidad predictiva. En un contexto regulatorio en constante evolución, como la Ley REDEC y la Ley de Finanzas Abiertas, el desarrollo del modelo posiciona a la empresa para aprovechar las oportunidades emergentes y adaptarse a nuevos desafíos en la gestión del crédito.

Bibliografía

Diario Financiero Chile . (10 de Agosto de 2023). Chile registró una mayor concentración del sector retail durante 2022. https://www.eleconomista.com.mx/empresas/Chile-registro-una-mayor-concentracion-del-sector-retail-durante-2022-20230810-0132.html

Ortiz, C. (21 de Noviembre de 2019). El retail, uno de los mayores empleadores del país y foco del conflicto social, en sus días cruciales. https://www.elmostrador.cl/noticias/2019/11/21/el-retail-en-dias-cruciales-uno-de-los-mayores-empleadores-del-pais-y-foco-del-conflicto-social/

Palacios, H. (25 de Mayo de 2022). Una mirada a la industria del retail en Chile. Clase Ejecutiva UC. https://www.claseejecutiva.uc.cl/blog/articulos/industria-del-retail-en-chile/

Banco Central de Chile. (2021, octubre). *Sistema financiero en Chile: Lecciones de la historia reciente*. https://www.bcentral.cl/documents/33528/133323/dpe67.pdf/08b40379-9553-fac0-f077-8ad5083f6a7f?t=1655149238017

Asociación Retail Financiero [ARF]. (Noviembre de 2023). Compendio Estadístico Mensual de la Industria del Crédito. https://retailfinanciero.org/wp-content/uploads/2024/03/Reporte-CMF-Bcentral-Sep-2023_vnov.pdf

CIEDESS. (2023). *Sube la morosidad en tarjetas de crédito de la banca y el retail y supera niveles pre pandemia*. Recuperado de https://www.ciedess.cl/601/w3-article-11382.html

Zack, M., & Newton, R. (2020). Managing risks associated with open loop payment systems. *Journal of Financial Risk Management, 11*(3), 245-260. https://doi.org/10.4236/jfrm.2020.113014

Visa. (2022). Four ways open loop systems make transit better for everyone. Recuperado de https://usa.visa.com/visa-everywhere/blog/bdp/2022/02/08/four-ways-open-1644350720138.html

CMF Educa (s.f.). ¿Qué es la CMF?. Recuperado el 27 de Marzo de 2024 de https://www.cmfchile.cl/educa/621/w3-article-49538.html

Comisión para el Mercado Financiero [CMF]. (Noviembre de 2021). CMF publica normativa que perfecciona la información de deudores del sistema financiero. https://www.cmfchile.cl/portal/prensa/615/w3-article-49336.html

Chakrabarty, N., & Biswas, S. (2018). *A Statistical Approach to Adult Census Income Level Prediction*. arXiv. https://doi.org/10.48550/arXiv.1810.10076

JPMorgan Chase Institute. (2018). Estimating family income from administrative banking data: A machine learning approach. JPMorgan Chase Institute. Recuperado de *https://www.jpmorganchase.com/content/dam/jpmc/jpmorgan-chase-and-co/institute/pdf/institute-estimating-family-income-report.pdf*

Parr, T., & Howard, J. (2020). How to explain gradient boosting. Explained.ai. Recuperado de https://explained.ai/gradient-boosting/

Doe, J., & Smith, L. (2018). Comparative Analysis of Internal and External Estimation Models. *International Journal of Finance, 22*(4), 456-478.

Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research, and Evaluation, 15*(12), 1-9.

Iglesias, P., Noguera, J., & Núñez, V. (2018). Data Transformation and its Impact on Machine Learning Model Performance. *Journal of Data Science, 16*(1), 45-59.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 4-37. https://doi.org/10.1109/34.824819

Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature, 55*(3), 789-865. https://doi.org/10.1257/jel.2016099

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880), 1164-1165. https://doi.org/10.1126/science.1154094

Instituto Nacional de Estadísticas. (2022). *Síntesis de Resultados Encuesta Suplementaria de Ingresos (ESI) 2022*. https://www.ine.gob.cl/docs/default-source/encuesta-suplementaria-de-ingresos/publicaciones-y-anuarios/síntesis-de-resultados/2022/síntesis-nacional-esi-2022.pdf?sfvrsn=529e421c_4

Instituto Nacional de Estadísticas. (n.d.). *Calculadora IPC*. Recuperado de https://calculadoraipc.ine.cl

Unholster & La Tercera. (2024). *Pensiones: edad de jubilación promedio cae casi tres años en las últimas cuatro décadas*. Recuperado de https://www.latercera.com/pulso/noticia/pensiones-edad-de-jubilacion-promedio-cae-casi-tres-anos-en-las-ultimas-cuatro-decadas/RN2OUL6QRFE7DPSE4PIOTQHTF4/

Rubin, D. B. (2004). Multiple Imputation for Nonresponse in Surveys. Wiley-Interscience.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.

OECD. (2023). *Disposable income by age group in Chile*. OECD Data Explorer. Recuperado de https://data-explorer.oecd.org/vis?df[ds]=DisseminateFinalDMZ&df[id]=DSD_WISE_IDD@DF_IDD&df[ag]=OECD.WISE.INE&dq=CHL.A.INC_DISP.MEAN..Y_GE76+Y66T75+Y_GT65+Y51T65+Y41T50+Y26T40+Y18T65+Y18T25+Y_LT18+_T.METH2012.D_CUR.&pd=2022,2023&to[TIME_PERIOD]=false&vw=tb.

Instituto Nacional de Estadísticas (INE). (2023). *Género y desigualdad de ingresos en Chile*. Recuperado de https://www.ine.gob.cl/docs/default-source/prensa-y-comunicacion/g%C3%A9nero-y-desigualdad-de-ingresos-en-chile.pdf?sfvrsn=64af7d60_2.

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

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146-3154.

Valenzuela Aros, P. (2016). *Determinantes y brecha de inclusión financiera en Chile y América Latina y el Caribe*. Tesis de Pregrado, Universidad de Chile. Recuperado de https://repositorio.uchile.cl/handle/2250/138958.

Ministerio de Salud. (2020). Se promulga nueva Ley Cenabast que permitirá disponer de remedios más baratos. Recuperado de https://www.minsal.cl/se-promulga-nueva-ley-cenabast-que-permitira-disponer-de-remedios-mas-baratos/

Palacios, A. (2020). Inclusión financiera y acceso al crédito en Chile: Un análisis del sector retail. *Revista de Economía y Sociedad, 15*(2), 45-60.

CMF Chile. (2024). *Normativa que regula el Sistema de Finanzas Abiertas en el marco de la Ley Fintech*. Recuperado de https://www.cmfchile.cl/portal/prensa/615/w3-article