Escuela de Administración

Determinar la viabilidad de implementar un modelo predictivo de recuperación de créditos para disminuir el índice de cartera vencida en el proceso de gestión y recuperación de cartera en una entidad financiera.

Proyecto Empresarial – Trabajo de Grado

Autor
Ana Solanyi Gamba Sotelo

Bogotá

2024

Escuela de Administración

Determinar la viabilidad de implementar un modelo predictivo de recuperación de créditos para disminuir el índice de cartera vencida en el proceso de gestión y recuperación de cartera en una entidad financiera.

Proyecto Empresarial – Trabajo de Grado

Autor
Ana Solanyi Gamba Sotelo

Tutor
Jesús Enrique Molina Muñoz

Master en Business Analytics

Escuela de Administración

22 de octubre de 2024

Bogotá, Colombia

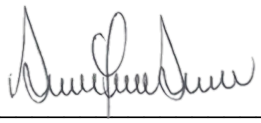
2024

Contenido

Preliminares

Declaración de originalidad y autonomía

Declaro bajo la gravedad del juramento, que he escrito el presente Proyecto Aplicado Empresarial (PAE), en la propuesta de solución a una problemática en el campo de conocimientos del programa de Maestría por nuestra propia cuenta y que, por lo tanto, su contenido es original. Declaro que he indicado clara y precisamente todas las fuentes directas e indirectas de información y que este PAE no ha sido entregado a ninguna otra institución con fines de calificación o publicación.

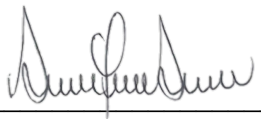


Ana Solanyi Gamba Sotelo

Firmado en Bogotá, D.C. el 9 de junio de 2024

Declaración de exoneración de responsabilidad

Declaro que la responsabilidad intelectual del presente trabajo es exclusivamente de su autor. La Universidad del Rosario no se hace responsable de contenidos, opiniones o ideologías expresadas total o parcialmente en él.



Ana Solanyi Gamba Sotelo

Firmado en Bogotá, D.C. el 09 de junio de 2024

Agradecimientos (Opcional)

Gracias a Dios por darme la oportunidad de poder cursar esta maestría, gracias a mi empresa por ser el patrocinador oficial, gracias a mi esposo, a Tony y Blu, por ser mi compañía siempre, gracias a mi mamá que desde el cielo me guía y acompaña.

Gracias a mi director de proyecto de grado y a mis profesores por su conocimiento y acompañamiento.

Ana Solanyi Gamba Sotelo

Lista de figuras

Lista de tablas

Abreviaturas

RC	Riesgo de crédito
SIAR	Sistema integral de administración de riesgo
EFoe	Entidad Financiera objeto de estudio
BA	Business Analytics,

Resumen Ejecutivo

El riesgo de crédito se entiende como la posibilidad de que una entidad financiera incurra en pérdidas y disminuya el valor de sus activos como consecuencia de que un deudor o contraparte incumpla sus obligaciones. (Superintendencia de Industria y Comercio, 2023) Poder acercarse o poder predecir esta probabilidad puede llegar a ser un factor distintivo en una etapa del ciclo del crédito como: la gestión de recuperación de cartera.

Por lo anterior este proyecto empresarial tiene como objetivo principal determinar la viabilidad de implementar herramientas de BA para disminuir el índice de cartera vencida en el proceso de gestión y recuperación de cartera de exempleados en la EFoe, a través de diferentes herramientas de BA como son, inteligencia de negocios, visualización de datos, gestión ágil de proyectos, ética y seguridad de datos, modelos estadísticos para la toma de decisiones, modelos de scoring, analytics life cycle management, gobernanza de sistemas de información y análisis de riesgos.

A su vez este proyecto de BA implica etapas fundamentales como el entendimiento del negocio, entendimiento y preparación de los datos, modelado, evaluación y despliegue.

Se realizará un modelo de scoring para recuperación de créditos a partir de la metodología logit, regresión logística y otros de machine learning, para determinar qué tan viable resulta para la organización, impulsando la eficiencia operativa y obteniendo una ventaja competitiva en un entorno empresarial en constante evolución, a su vez busca obtener información precisa sobre las posibilidades de incumplimiento y tomar decisiones asertivas alineadas con la estrategia empresarial, esto con el fin de conducir a la EFoe a mantener un indicador de calidad de cartera alineado al sector, para evitar o reducir el incumplimiento en

el pago de los créditos aprobados y plantear estrategias para gestionar de manera diferencial a los deudores, fundamentadas en la recopilación, procesamiento y análisis de datos.

Palabras clave

Análisis de negocios, modelos estadísticos para la toma de decisiones, modelos de scoring, gestión de riesgo de crédito.

Abstract

Credit risk is understood as the possibility that a financial entity may incur losses and decrease the value of its assets as a result of a debtor or counterparty failing to fulfill their obligations (Superintendencia de Industria y Comercio, 2023). Being able to approach or predict this probability can become a distinctive factor in a stage of the credit cycle such as portfolio recovery management.

Therefore, this business project aims primarily to determine the feasibility of implementing BA tools to reduce the delinquency rate in the management and recovery process of former employees portfolios at EFoe, through various BA tools such as business intelligence, data visualization, agile project management, ethics and data security, statistical models for decision-making, scoring models, analytics life cycle management, information systems governance, and risk analysis.

Moreover, this BA project involves key stages such as understanding the business, data understanding and preparation, modeling, evaluation, and deployment. A scoring model will be developed for credit recovery using methodologies like logit, logistic regression, and other machine learning techniques, to assess the organization's viability, enhance operational efficiency, and gain a competitive advantage in a continually evolving business environment. It also aims to obtain precise information on default probabilities and make assertive decisions aligned with the business strategy, with the goal of guiding EFoe to maintain a portfolio quality indicator aligned with the sector, to prevent or reduce default on approved credits and propose differential strategies for managing debtors, based on data collection, processing, and analysis.

Keywords

Business analysis, statistical models for decision-making, scoring models, credit risk management.

1 Introducción

Debido a las tasas de interés altas, que buscan presionar la inflación a la baja, se ha vuelto más difícil para los colombianos cumplir con sus deudas. Según las cifras del sistema de la Superfinanciera, a febrero de 2024, los consumidores mantienen la demora con los pagos de sus responsabilidades crediticias que se ven afectadas por la desaceleración económica.

Y es que, la cartera vencida de febrero quedó en $35 billones, cifra que deja un sin sabor, pues, aunque en el comparativo mensual logró reducir $192.552 millones, en el anual mostró un crecimiento de $7,4 billones. Este último dato refleja una variación real anual de 17,81% (esta variación descuenta el dato de inflación), pues en febrero de 2023 estaba en $27,5 billones. El total de la cartera bruta vencida registra un aumento anual de $11.346 millones que representa un incremento en 5,62%, a su vez la cartera de vivienda vencida también registró un aumento significativo de $9.768 millones que representa un aumento del 1.35% anual. Este contexto económico adverso tiene consecuencias palpables, no solo para los consumidores sino también para las entidades financieras, que se ven obligadas a ser más rigurosas en la concesión de créditos y a su vez la capacidad de pago de los deudores se ve comprometida, ya que tanto las personas naturales como las empresas enfrentan costos inflacionarios, llevando a un consumo básico y la postergación de obligaciones financieras. Este escenario sugiere la necesidad de medidas cautelares para mitigar el impacto en la economía nacional. (La República, 2024).

Por su parte la EFoe en 2023 y lo corrido del 2024 no ha sido ajena a esta situación y ha enfrentado grandes desafíos en cuanto a la recuperación y gestión de cartera; el indicador de calidad de cartera vencida de la cartera de exempleados ha experimentado cambios

significativos en su resultado, este indicador mide la mora en las carteras de vivienda, consumo y vivienda + consumo; este indicador es comparado con el promedio del total de los establecimientos de crédito del país, durante el último semestre de 2023 y lo corrido del 2024 el indicador de calidad de cartera total vivienda alcanzó un valor máximo de 6.49% mientras el resultado máximo de las entidades de crédito fue del 3.14%, superar este indicador es una alerta para la Efoe en donde aplicar nuevas estrategias como un modelo predictivo de scoring que determine el riesgo que un deudor incumpla en el pago de su próxima cuota, puede ayudar a tomar decisiones en la gestión diferencial de recuperación de cartera.

Adicionalmente a las consideraciones mencionadas anteriormente se suma un factor que condiciona la recuperación de cartera y es la expedición de la Ley 2300 de 2023 por medio de la cual se establecen medidas que protejan el derecho a la intimidad de los consumidores, esta ley debe ser implementada por los las entidades vigiladas por la Superintendencia Financiera y todas las personas naturales y jurídicas que adelanten gestiones de cobranzas de forma directa, esta ley busca proteger el derecho a la intimidad de los consumidores y regular la cantidad de mensajes publicitarios, canales y horarios en los que las personas pueden ser contactadas para realizar gestión de cobro o recibir información de carácter comercial. Ley 2300 de 2023 (Congreso de la República de Colombia, 2023). Esta ley condiciona y de alguna manera restringe la manera de realizar actividades de cobranza y hace que la EFoe deba incorporar nuevas estrategias que permita alcanzar una recuperación de cartera óptima, pues al disminuir los canales autorizados o al no tener la autorización formal de los canales por los cuales autorizan a ser contactados (llamadas, WhatsApp, correo electrónico, mensaje de texto), esto sumado a restricciones de días y horarios, a la limitación de un solo contacto directo semanal y al no poder preguntar por la razones del no pago, hacen que la gestión de

cartera se restrinja y se deben buscar estrategias más óptimas que permita la correcta gestión de la recuperación de cartera

"Adicionalmente la normativa en Colombia exige ciertas condiciones para el Sistema Integral de Administración de Riesgo SIAR, este debe contener políticas y procedimientos claros y precisos que definan los criterios y la forma mediante la cual la entidad evalúa, asume, califica, controla y cubre su riesgo crediticio. Para ello, los órganos de dirección, administración y control de las entidades deben adoptar políticas y mecanismos especiales para la adecuada administración del riesgo crediticio, no sólo desde la perspectiva de su cubrimiento a través de un sistema de provisiones, sino también a través de la administración del proceso de otorgamiento de créditos y permanente seguimiento", (Superintendencia de Industria y Comercio, 2023), de éstos es claro que la EFoe cuenta con un SIAR (Sistema integral de administración de Riesgo) establecido y maduro.

Teniendo en cuenta lo anterior en el ámbito financiero, la gestión eficiente de las carteras de créditos es de vital importancia para garantizar la estabilidad y el crecimiento sostenible. De esta manera la EFoe, reconocida como una entidad financiera líder en el país, reconoce y entiende la importancia de implementar nuevas acciones que conduzca al cumplimiento de sus objetivos e indicadores de calidad de cartera; es por esto que contar con el planteamiento y diseño de un modelo de scoring de recuperación de créditos, que permita lograr una medición precisa, real, eficiente de la probabilidad de incumplimiento en el pago en los créditos de la cartera de empleados para enfocar esa gestión en clientes críticos, disminuir tiempos y optimizar el proceso. Adicionalmente, el proyecto desarrollará un ambiente tecnológico suficiente, seguro y con los protocolos necesarios que respalden la

implementación y el monitoreo continuo de las herramientas de BA. De esta manera los usuarios gestores e interesados podrán acceder a la información actualizada sobre los valores de probabilidad de incumplimiento.

Lograr la consecución del objetivo de este proyecto y su correcta implementación suministrará a la EFoe una herramienta de Business Analytics completa para el control de riesgo de crédito de la cartera de empleados y exempleados. Esto, a su vez, fortalecerá el enfoque financiero y mejorará su capacidad para ofrecer productos de crédito, confiables y sostenibles.

El equipo responsable del proyecto cuenta con experiencia en gestión de riesgos y análisis de datos, asegurando una implementación exitosa y una transición fluida hacia el nuevo modelo de gestión y predicción, si se demuestra viable.

2 Objetivos

2.1 Objetivo general

Determinar la viabilidad de implementar un modelo predictivo de recuperación de créditos para disminuir el índice de cartera vencida en el proceso de gestión y recuperación de cartera en una entidad financiera

2.2 Objetivos específicos

- Analizar a profundidad el contexto y los objetivos específicos de La EFoe para identificar las necesidades y oportunidades claves para la gestión de recuperación de cartera de exempleados, alineando el proyecto con la estrategia e indicadores del negocio.
- Diseñar modelos predictivos que permitan una evaluación precisa y en tiempo real de la probabilidad de riesgo de no pago de los créditos, facilitando la toma de decisiones informadas y oportunas, en el proceso de gestión y recuperación de cartera.
- Determinar la viabilidad de implementación en la EFoe de las herramientas que resulten necesarias para determinar la probabilidad de incumplimiento de pago en los créditos de la cartera de exempleados de la EFoe y proporcionar conclusiones y recomendaciones.

3 Alcance

El alcance del proyecto inicia desde la comprensión y el conocimiento del negocio hasta el estudio de viabilidad de implementar un modelo predictivo de scoring de recuperación de créditos con el objetivo de identificar el riesgo de incumplimiento del pago de créditos de la cartera de exempleados de la EFoc. Para lograrlo, se seguirán los pasos correspondientes para llevar a cabo un análisis efectivo. Se espera desarrollar un modelo que permita a la EFoe identificar de manera anticipada la posibilidad de incumplimiento en el pago de un crédito, lo cual servirá como herramienta de para tomar decisiones sobre la gestión de recuperación de cartera y segmentaciones de deudores de la cartera de exempleados.

Es importante tener en cuenta que el alcance del proyecto puede ajustarse durante su ejecución, con el fin de adaptarse a las necesidades y requerimientos específicos de la compañía y garantizar la efectividad del modelo en la gestión y control de las carteras mencionadas.

4 Marco Teórico

Con el fin de lograr cumplir los objetivos planteados en este proyecto aplicado, se han examinado las siguientes teorías, modelos y técnicas analíticas para la predicción de la variable el deudor pagara o no la siguiente cuota de su crédito.

4.1 Riesgo de Crédito.

El riesgo de crédito, también denominado riesgo crediticio, se define como la probabilidad de que un prestatario incumpla con sus obligaciones de pago o, de manera alternativa, que una entidad financiera no logre recuperar el dinero prestado. En este sentido, el riesgo de crédito representa una variable crucial en el ámbito económico, tanto para las instituciones financieras como para los prestatarios, dado que incide directamente en el costo y las condiciones de los préstamos. (Banco Santander, 2023).

4.2 Gestión del Riesgo de Crédito.

En términos prácticos, la gestión del riesgo de crédito implica implementar estrategias para mitigar posibles pérdidas, considerando la solvencia de la entidad y las reservas disponibles para enfrentar tales contingencias. Este tipo de riesgo se relaciona de manera específica con los problemas financieros de la entidad prestataria, diferenciándose del riesgo de mercado, el cual, como se analizará posteriormente, tiene una dimensión más sistemática y afecta a los mercados en su conjunto. (Banco Santander, 2023).

4.3 Cartera de créditos.

Monto total de los préstamos que hacen todos los deudores financieros. (Superintendencia Financiera de Colombia, 2023).

4.3.1 Cartera vencida.

Monto de la cartera bruta que se encuentra en mora de pagos. (Superintendencia Financiera de Colombia, 2023).

4.3.2 Calidad de cartera.

Proporción de la cartera vencida sobre la cartera bruta expresada en porcentaje para indicar la morosidad de la misma. (Superintendencia Financiera de Colombia, 2023).

4.3.3 Indicador calidad de cartera total

Este indicador mide el total del saldo de la cartera que se encuentra en mora entre 31 y 540 días de mora. Formula

$$ICT = \frac{Saldo\ de\ la\ cartera\ total\ (vivienda + consumo) en\ mora\ entre\ 31\ y\ 540\ días}{Saldo\ total\ de\ la\ cartera\ (cartera\ vivienda\ y\ consumo\ al\ día\ y\ en\ mora)}$$

4.4 Modelo de Regresión Logística.

Este modelo se enmarca en la familia de los modelos de respuesta cualitativa, donde la variable dependiente es binaria, tomando únicamente dos valores: 1, si el objeto de estudio pertenece a una condición específica (éxito), o 0, si pertenece a la condición opuesta (fracaso). Esta estructura simplifica el cálculo de la probabilidad de que un cliente pertenezca a uno de los grupos predefinidos, como no pagador o pagador.

Una de las ventajas destacadas de este modelo radica en que mantiene la variable explicada en un rango entre cero y uno, lo que lo hace idóneo para medir la probabilidad de incumplimiento. La elección de la logística como una opción válida se atribuye al físico y estadístico Joseph Berkson, quien en 1944 propuso su utilización, acuñando el término 'logit' (Cramer, 2003).

Esta probabilidad se formula mediante la ecuación ln(P/(1-P)), donde P representa la probabilidad de que ocurra un evento específico. La ventaja distintiva de la regresión logística radica en su facilidad de interpretación, ya que modela directamente la probabilidad logística y proporciona un intervalo de confianza para el resultado. (Chiu, 2015).

4.5 Modelo de Scoring.

Un modelo de scoring (Mester, 1997) es un método de evaluar el riesgo de crédito de solicitudes de préstamos (scoring de admisión) o de préstamos ya concedidos anteriormente (scoring de comportamiento). El objetivo es aislar el efecto de una serie de características personales o propias del producto en la probabilidad de impago del cliente, utilizando datos históricos y técnicas estadísticas. El modelo da como resultado una puntuación o "score" que la Efoe puede utilizar para calificar al cliente y determinar el pago o impago. Para ello se utiliza información histórica de la entidad de créditos concedidos que han resultado en impago, créditos concedidos que han acabado satisfactoriamente y créditos no concedidos a los que se les estimará un comportamiento (técnica conocida como "reject inference").

Scoring para la recuperación de créditos, evalúa a los clientes según su probabilidad de retraso en los pagos. Esto facilita una mejor gestión de los clientes irregulares (aquellos con un retraso en algún pago de menos de 90 días) y permite tomar decisiones de recuperación basadas en el nivel de morosidad o irregularidad.

La tarjeta de puntuación ofrece una manera clara y accesible de presentar los resultados obtenidos a partir de la compleja metodología del scoring.

Weight of Evidence (WOE) El peso de la evidencia o WOE indica el poder predictivo de cada categoría, Nieto (2010). Mide la diferencia entre la proporción de buenos y malos en cada grupo.

Se define de la siguiente manera:

$$WOE_i = ln\left(\frac{\%Buenos_i}{\%Malos_i}\right)$$

Donde: $\%\, Buenos_i = \left(\frac{Buenos_i}{Total\ buenos}\right)$ y $\%\, Malos_i = \left(\frac{Malos_i}{Total\ malos}\right)$

El uso de WOES es particularmente adecuado para modelar utilizando regresión logística.

Information Value (IV) el Information Value es una medida muy usada en la construcción de scorecards. Con este estadístico se puede medir el poder de predicción de agrupar los atributos de una variable. Además, es un buen indicador a la hora de seleccionar variables para un modelo de regresión logística binario, cómo es el caso de un modelo de scoring. Se define así:

$$IV = \sum_{i=1}^{k} (\%Buenos_i - \%Malos_i) * WOE_i$$

Obtención de los scores y Scorecard. una vez que las variables han sido categorizadas, los WOE's han sido calculados y los betas del modelo han sido obtenidos, es posible reescalar las probabilidades estimadas en puntuaciones, de acuerdo a la siguiente fórmula (Nieto, 2010):

$$Score = \left(\frac{\text{Off set}}{n}\right) + \left(\frac{\text{Factor}}{\ln(2)}\right) * \left(\frac{\beta_0}{\text{n}} - \beta_{ij} * WOE_{ij}\right)$$

Donde i=variable, j=categoría, n es el número de variables off set 500, factor 20

4.6 Modelo Naive Bayes.

El algoritmo clasificador Naïve-Bayes (NBC), es un clasificador probabilístico simple con fuerte suposición de independencia. Aunque la suposición de la independencia de los

atributos es generalmente una suposición pobre y se viola a menudo para los conjuntos de datos verdaderos. A menudo proporciona una mejor precisión de clasificación en conjuntos de datos en tiempo real que cualquier otro clasificador. También requiere una pequeña cantidad de datos de entrenamiento. El clasificador Naïve-Bayes aprende de los datos de entrenamiento y luego predice la clase de la instancia de prueba con la mayor probabilidad posterior. También es útil para datos dimensionales altos ya que la probabilidad de cada atributo se estima independientemente (Chandra et al, 2007).

4.7 Modelo de máquinas de soporte vectorial

La máquina de vectores de soporte (SVM, por sus siglas en inglés) es un algoritmo avanzado de clasificación y regresión que se basa en principios de aprendizaje automático para maximizar la precisión de las predicciones mientras evita el sobreajuste de los datos. El proceso de SVM puede incluir una transformación no lineal opcional de los datos de entrenamiento, seguida por la formulación de ecuaciones de regresión en el espacio de datos transformado. Esta metodología permite separar las clases en problemas de clasificación (para objetivos categóricos) o ajustar valores en problemas de regresión (para objetivos continuos). (IBM, 2021)

4.8 Arboles de clasificación.

Los modelos de agrupación en clústeres según IBM (IBM, 2021) se especializan en agrupar registros con los mismos criterios y que una vez agrupados estos sean identificados mediante etiquetas, lo cual facilitaría la identificación de grupos con las mismas características. Estos modelos cluster (agrupan) según la cantidad que se requiera y/o según el mismo modelo identifique los grupos con mayor coincidencia, ya que, no es necesario

que se tenga un conocimiento completo de todos los criterios con los que cuentan los individuos analizados.

Es una técnica utilizada para organizar datos de manera jerárquica según características específicas. Funciona haciendo una serie de preguntas sobre cada dato para separarlo en diferentes grupos. Cada pregunta se basa en una característica del dato, como, por ejemplo, "¿Cuál es el endeudamiento?" o "¿tiene cuentas por cobrar?". A medida que se responden estas preguntas, el árbol de clasificación va dividiendo los datos en grupos más pequeños y específicos. Esto facilita la organización y el entendimiento de conjuntos de datos complejos, ayudando a clasificar datos, ingresos o cualquier tipo de información de manera ordenada y eficaz.

4.9 Random Forest.

Dentro del aprendizaje supervisado, el algoritmo Random Forest, propuesto por (Breiman 2001), se destaca como una técnica que construye múltiples árboles de decisión basados en un conjunto de datos de entrenamiento. Los resultados de estos árboles se combinan para formar un modelo final que presenta una mayor robustez en comparación con los resultados individuales de cada árbol. El proceso de construcción de un modelo Random Forest se realiza en dos etapas principales:

- Se genera un número significativo de árboles de decisión utilizando el conjunto de datos. Cada árbol se construye a partir de un subconjunto aleatorio de variables (predictores.
- Cada árbol se desarrolla hasta su máxima extensión sin poda.

Adicionalmente, cada árbol en el algoritmo Random Forest incluye un conjunto de observaciones aleatorias seleccionadas mediante la técnica de bootstrap. Este método estadístico permite obtener muestras de una población en las que una observación puede aparecer en más de una muestra. Las observaciones no utilizadas en la construcción de los árboles, conocidas como “out of the bag,” se emplean para validar el modelo. Finalmente, las predicciones de todos los árboles se consolidan en una salida final, a través de un proceso de ensamblaje que se realiza mediante una regla específica. Esta regla suele ser el promedio en el caso de salidas numéricas o el conteo de votos en el caso de salidas categóricas (Lizares, 2017).

4.10 Metodología CRISP-DM (Manual CRISP-DM de IBM SPSS Modeler, 2021).

4.10.1 Comprensión del negocio

Es la primera etapa del proceso y tal vez la más importante de todo el proceso. En esta etapa se debe llegar a la comprensión de cuál es el propósito y los requisitos del proyecto vistos desde una perspectiva empresarial. Adicionalmente se debe aprovechar al máximo la extracción de datos y es muy importante comprender de manera integral el problema al cual se espera dar una solución. Realizar estas actividades facilitará la recolección de datos y se podrá realizar una interpretación acertada de los resultados, identificar correctamente esas variables que pueden ser significativas en el proyecto, no redundar en información y no recolectar información que no genere ningún valor agregado pude ayudar al cumplimiento del entendimiento del negocio. Durante esta fase, es muy importante tener las habilidades de traducir el conocimiento empresarial adquirido en un problema de extracción de datos y en un plan preliminar con el objetivo de cumplir con los objetivos empresariales.

4.10.2 Comprensión de los datos

Esta segunda fase de la metodología CRISP-DM, busca hacer una captura inicial de los datos que se debe analizar para poder familiarizarse con los mismos, se deben identificar esos problemas de calidad que se estén presentando identificar que se puede extraer de estos datos y detectar otros conjuntos de datos que permitan formular hipótesis.

4.10.3 Preparación de datos

Después de identificar las fuentes iniciales de los datos y luego que estos han sido recolectados, se debe continuar con la limpieza de estos, esto con el fin de aumentar la calidad de los datos, se deben identificar los valores vacíos, duplicados, nulos, y definir qué técnica o técnicas se utilizarán para cumplir con esta limpieza.

Adicional a esto es importante realizar ingeniería de características partiendo de la información inicial recolectada a partir de esta se pueden construir variables muy significativas y de gran aporte para la construcción del modelo, por último, se debe realizar una integración de datos para crear nuevos o establecer una fuente de datos consolidada.

4.10.4 Modelado

En esta fase de CRISP-DM se seleccionan y aplican diferentes técnicas de modelado, las técnicas pueden ser variadas y cada una de ellas exige ciertas especificaciones, teniendo en cuenta lo anterior es importante mencionar que en esta fase es posible que se requieran ciclos adicionales de preparación de datos. Durante esta fase las principales actividades que se deben realizar son: selección de la técnica de modelado, diseño y separación de los datos que se usarán para el entrenamiento, testeo y validación, en esta fase también se debe construir

el modelo, por último, se debe evaluar el modelo basado en los objetivos de éxito determinar si se deben ajustar las técnicas utilizadas para así lograr una mayor calidad en el resultado.

4.10.5 Evaluación

Esta etapa tiene lugar una vez se ha construido el o los modelos y se elige el que presente mayor calidad desde los datos y objetivos del negocio, el resultado final de esta fase es decidir si se aprueba la utilización o no de los resultados. Las principales actividades que incluye esta fase son: evaluación de los resultados se centra en la validación del cumplimiento de los objetivos del negocio, determinar si existen razones que hagan el modelo no eficiente, también si el alcance, tiempo y costos lo permiten; revisión del proceso reevaluar los pasos ejecutados y determinar si se han pasado por alto factores importantes; decisión sobre los siguientes pasos según las conclusiones encontradas ya sea continuar con la implementación para iniciar la operación, hacer nuevas iteraciones o dar por finalizado el proyecto.

4.10.6 Implementación

En esta fase, y una vez que el modelo ha sido construido y validado, se deben realizar ciertas actividades como: planificar la estrategia de implementación y los pasos a que se requieran (como, quien, cuando), se debe planificar cómo se va a monitorear y mantener el modelo, realizar un informe final con todos los documentos necesarios y relevantes del proyecto, por último, definir las lecciones aprendidas.

4.11 Evaluación de los Modelos.

Figura 1 Cuadro Comparativo de Modelos de Aprendizaje Automático

Modelo	Simplicidad	Adaptabilidad	Interpretación	Mantenimiento
Random Forest	Bajo	Alto	Alto	Moderado
Modelo de Scoring de Crédito	Moderado	Bajo	Alto	Bajo
Árbol de Clasificación	Moderado	Moderado	Moderado	Bajo
Máquina de Soporte Vectorial (SVM)	Bajo	Alto	Bajo	Alto
Modelo de Regresión Logística	Alto	Moderado	Alto	Bajo
Modelo de Naive Bayes	Alto	Moderado	Moderado	Bajo

Nota. Cuadro Comparativo de Modelos de Aprendizaje Automático.
Fuente: (Christopher M. Bishop)

4.12 Métricas para la Evaluación de los Modelos. (Murphy, K. P. 2012).

- Accuracy (Exactitud): Proporción de predicciones correctas en relación con el total de predicciones
- 95% CI (Intervalo de Confianza al 95%): Indica el rango en el cual es probable que se encuentre el verdadero valor de exactitud.
- No Information Rate (Tasa de No Información): Es el rendimiento que se lograría si simplemente predices siempre la clase más frecuente.
- P-Value [Acc > NIR] (Valor p para Exactitud > Tasa de No Información): Un valor p bajo indica que la exactitud del modelo es significativamente diferente de la tasa de no información.
- Kappa: El coeficiente kappa, que mide la concordancia más allá del azar. Valores más cercanos a 1 indican una concordancia más fuerte.
- Mcnemar's Test P-Value (Valor p del test de McNemar): este test evalúa si hay una diferencia significativa en las tasas de error entre dos modelos o clasificadores

- Sensitivity (Sensibilidad): La proporción de casos positivos que fueron correctamente identificados por el modelo.
- Specificity (Especificidad): La proporción de casos negativos que fueron correctamente identificados por el modelo.
- Pos Prep Value (Valor Predictivo Positivo): La proporción de predicciones positivas correctas en relación con el total de predicciones positivas.
- Neg Pred Value (Valor Predictivo Negativo): La proporción de predicciones negativas correctas en relación con el total de predicciones negativas.
- Prevalence (Prevalencia): La proporción de la clase positiva en los datos
- Detection Rate (Tasa de Detección): La proporción de casos positivos que fueron correctamente identificados por el modelo.
- Detection Prevalence (Prevalencia de Detección): La proporción de casos que fueron clasificados como positivos por el modelo.
- Balanced Accuracy (Exactitud Equilibrada): El promedio de la sensibilidad y la especificidad. Es una medida útil cuando las clases están desbalanceadas.

5 Metodología

5.1 Metodología del proyecto CRISP-DM

Este proyecto empresarial se desarrollará siguiendo la metodología CRISP-DM, la cual se encuentra estructurada en 6 fases, abordadas de manera secuencial, cabe resaltar que la metodología puede ser aplicada bidireccionalmente, lo que significa que se puede retroceder o volver a pasar por alguna etapa, cuando este sea considerado, ya sea por redefiniciones o necesidades del proyecto, se trabajará el proyecto acorde a estas fases, explicando cómo se realizará la aplicación a las fuentes de información, al planteamiento del problema y a la solución buscada (Manual CRISP-DM de IBM SPSS Modeler, 2021).

Figura 2 Metodología CRISP -DM

Nota. La figura muestra las fases de la metodología CRISP-DM. Fuente: Instituto de Ingeniería del Conocimiento (2021).

5.2 Metodologías ágiles

Usando esta metodología se ha realizado el cronograma de trabajo, registrado los sprint y respectivos entregables. Adicionalmente se determinan factores importantes de la organización a través de las historias de usuario y su priorización, de igual forma se realizará la matriz DOFA que permitirá conocer más a profundidad sobre el contexto del negocio y su problemática.

6 Cronograma

A continuación, se muestra el cronograma establecido para el proyecto, el cual es planteado bajo metodología Crisp – Dm metodologías ágiles y sprints:

Figura 3 Cronograma

CRONOGRAMA				
Tareas	Sprint	Inicio	Fin	Estado
Definición de resumen ejecutivo, introducción, objetivos y alcance Definición del cronograma Identificación fuentes de información Diseño de metodología a aplicar	Anteproyecto	*05/may/2023*	*16/jun/2023*	*Finalizado*
Descripción de la situación organizacional y problemática	Desarrollar matriz Foda, identificación de factores estratégicos y problematica. Estado actual, ética, gobernanza de datos para el proyecto	*01/jul/2023*	*20/08/2023*	*Finalizado*
Revisión y retroalimentación director de proyecto	Revisión y ajustes	*15/ago/2023*	*15/ago/2023*	*Finalizado*
Construcción de la base de datos	Revisión de variables disponibles, recopilación y construcción de la base de datos	*21/ago/2023*	*10/oct/2023*	*Finalizado*
Aplicación metodología Crisp-DM	Comprensión de los datos, análisis en la calidad y dimensión de los datos, exploración de los datos, analytics lifecycle management	*10/oct/2023*	*31/nov/2023*	*Finalizado*
Análisis y depuración de las bases de datos	Limpieza de datos	*01/nov/2023*	*10/nov/2023*	*Finalizado*
Visualización actual	Visualización de cifras de cartera de la cartera de exempleados actual	*11/nov/2023*	*20/nov/2023*	*Finalizado*
Retroalimentación por director	Identificación de oportunidades de mejora	*21/nov/2023*	*21/nov/2023*	*Finalizado*
Revisión de bases de datos	Inclusión y eliminación de variables	*22/nov/2023*	*24/nov/2023*	*Finalizado*
Realizar modelo de regresión logística	Modelo I	*25/nov/2023*	*30/nov/2023*	*Finalizado*
Retroalimentación por director	Revisión de resultados	*01/dic/2023*	*01/dic/2023*	*Finalizado*
Entregable de trabajo de grado II	Entrega documento trabajo de grado II	*15/dic/2023*	*15/dic/2023*	*Finalizado*
Presentación trabajo de grado II	Realizar presentación trabajo de grado II	*16/dic/2023*	*16/dic/2023*	*Finalizado*
Recolección de datos adicionales	Recolección de variables adicionales	*10/feb/2024*	*20/feb/2024*	*Finalizado*
Realizar modelo predictivo de scoring de recuperación de créditos	Modelo II	*01/abr/2024*	*21/may/2024*	*Finalizado*
Incluir temas adicionales	Temario III semestre	*10/abr/2024*	*23/may/2024*	*Finalizado*
Realizar conclusiones y recomendaciones	Incluir conclusiones y recomendaciones	*24/may/2024*	*31/may/2024*	*Finalizado*
Realizar documento final	Terminar documento del proyecto	*01/may/2024*	*31/may/2024*	*Finalizado*
Revisión final	Realizar revisión final con el director del proyecto	*05/jun/2024*	*05/jun/2024*	*Finalizado*
Sustentación	Realizar sustentación del proyecto		*jul/2024*	

Nota. Cronograma proyecto empresarial. Fuente: elaboración propia.

7 Descripción de la Situación organizacional

La EFoe tiene como objeto principal 7 actividades las cuales son el core de su negocio: redescuento de créditos, recibo de depósitos de entidades públicas, captación de ahorro, operaciones de crédito externo, servicios de asistencia técnica, estructuración y consultoría de proyectos, administración de títulos y emisión de avales y garantías.

La EFoe es una empresa familiarmente responsable (EFR), durante sus 34 años de existencia se ha caracterizado por brindar beneficios más allá de los determinados por la normatividad laboral vigente, dentro de su convención colectiva de trabajo se estipula que uno de estos beneficios es el otorgamientos de créditos a empleados que cumplan cierta antigüedad en la entidad, las líneas de créditos que se otorgaran a los empleados son: crédito de vivienda, libre destino, estudió, vehículo y/o calamidad doméstica. El área responsable de realizar la recepción y evaluación de documentos en el área de Talento Humano, después de esto la solicitud de crédito es llevada a un comité donde se aprueba o no el crédito.

Una vez el crédito es desembolsado la operación de crédito pasa a ser responsabilidad de la Dirección de Cartera quien será el área encargada de gestionar la recuperación de cartera durante toda la vida del crédito y custodiar las garantías del crédito, el objetivo principal de la dirección de cartera en la EFoe es: Administrar las diferentes carteras que maneja la EFoe y sus respectivas garantías, gestionando oportunamente los procedimientos establecidos, así como la normatividad vigente, para lograr su recuperación dentro de los plazos definidos, asegurando la integridad, confidencialidad y disponibilidad de la información y custodiando las garantías desde el momento del desembolso hasta su cancelación final por parte del deudor. Teniendo en cuenta lo anterior es evidente que la recuperación de cartera es el core de esta dirección y su principal objetivo.

Dentro de las diferentes carteras que debe gestionar la EFoe se encuentran las carteras de empleados y exempleados las cuales es importante definir.

- **Cartera Empleados:** Cartera originada por los préstamos otorgados a los empleados de la EFoe.
- **Cartera Exempleados:** Cartera originada por los préstamos otorgados a los empleados de la EFoe, quienes ya no laboran en la Entidad.

Continuando con la explicación del negocio la cartera de empleados es una cartera que no requiere gestión de recuperación de cartera pues una vez aprobado y desembolsado el crédito, este se cobra a través de descuento automático quincenal por nómina, es decir que no existe ninguna posibilidad que un deudor de la cartera de empleados entre en mora pues en el estudio de viabilidad del crédito fue validado que la cuota total mensual a pagar pudiera ser cubierta por el salario del empleado.

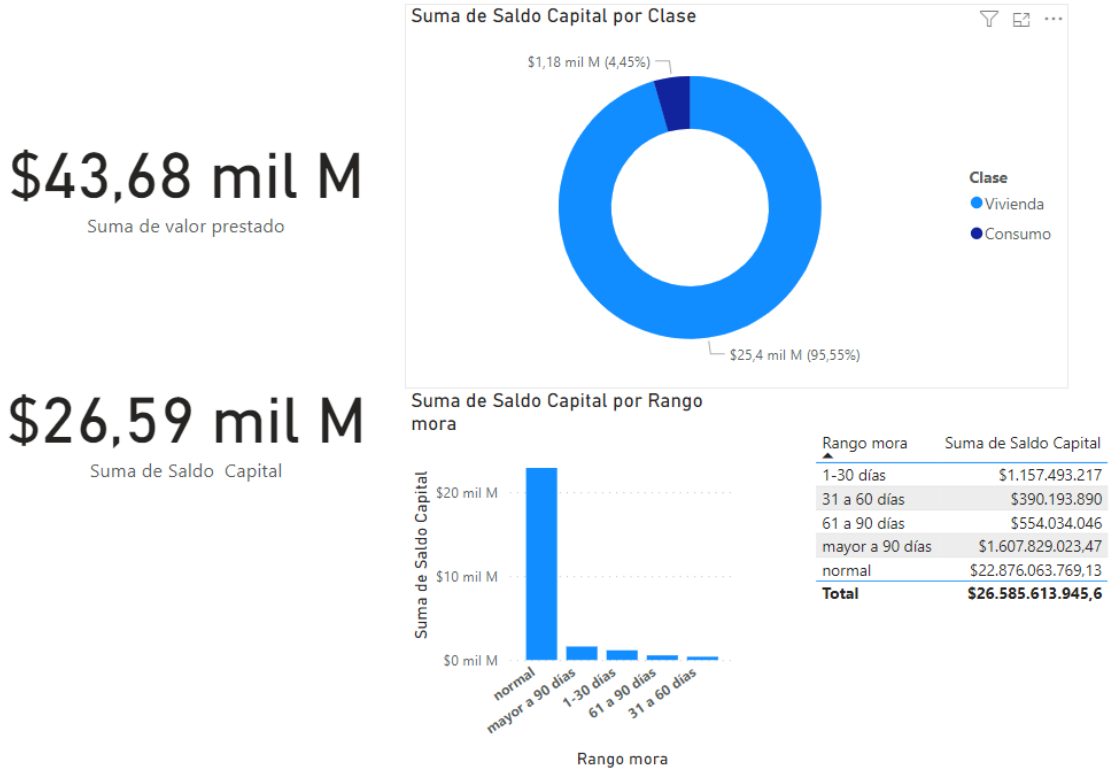
Ahora bien y en donde está el objeto de estudio de este proyecto es en la cartera de exempleados, que como se mencionó anteriormente ocurre cuando un empleado deja de estar vinculado a la EFoe por principalmente 4 razones (renuncia voluntaria del trabajador, despido por justa causa, despido sin justa causa, pensión por jubilación o enfermedad), una vez el empleado se convierte en exempleado ya no cuenta con el ingreso de su salario en la EFoe y es en este momento donde inicia la gestión de recuperación de cartera.

Desde el 2020 la Dirección de Cartera se consolidó como una única dirección y entregó las funciones de estudio de crédito a la Dirección de crédito, desde ese entonces se han realizado múltiples capacitaciones y diplomados en gestión eficiente de cartera, y desde el 2022 se viene elaborando mensualmente un plan de gestión de cartera que analiza el resultado del cierre de un mes determinado y el mes inmediatamente anterior y categoriza el

monto total de la cartera que incurrió en mora para las franjas de cartera al día, cartera en mora de 1 a 30 días, cartera en mora de 31 a 60 días, cartera en mora de 61 a 90 días y mora superior a 91 días. Adicional a esto de manera frecuente se actualiza la matriz FODA para analizar las fortalezas, oportunidades, debilidades y amenazas que se pueden presentar durante la gestión de cobranza, continuando con el plan se establecen unos objetivos de qué cartera se debe conservar el día, que franjas de mora se deben mantener y si es procedente que deudores específicos van a deteriorarse rodando a una franja de mora superior y cuáles van a mejorar rodando a una franja de mora inferior, como última etapa del plan de gestión de cobranza se establecen las actividades y estrategias a implementar para construir esos objetivos planteados. Este ejercicio se hace todos los meses y se gestiona a todos los deudores que presenten desde un día de mora, esta situación es desgastante y muchas veces poco eficiente, incorporar un modelo de scoring de recuperación de créditos, que muestre la probabilidad que un deudor incurra en mora puede ayudar a gestionar realmente a los deudores clave y establecer estrategias y tomar decisiones sobre la gestión de la cartera, mejorando el indicador de calidad de cartera y evitando altos valores en provisiones por créditos de exempleados deteriorados o incobrables.

7.1 Visualización situación actual

Figura 4 Visualización situación actual



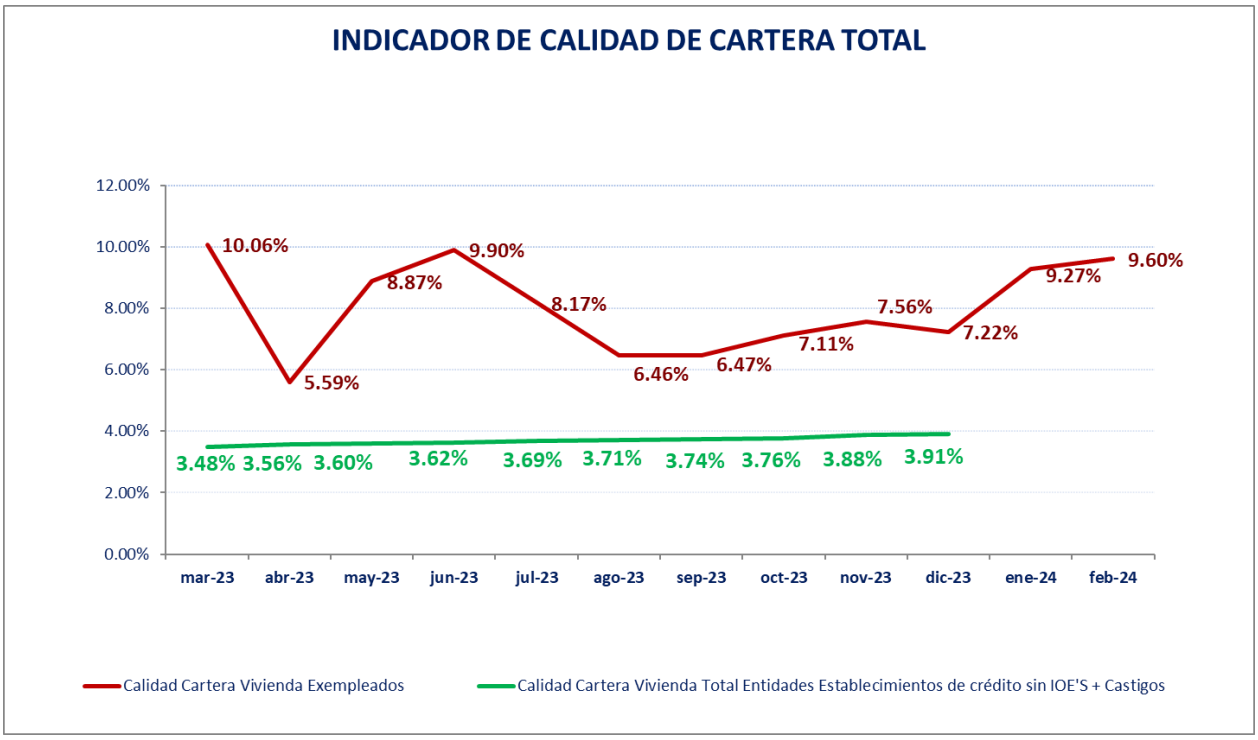
Nota. Visualización de datos cartera de exempleados actual. Fuente: elaboración propia.

8 Descripción de la problemática empresarial y método a aplicar para su solución

Para iniciar con el desarrollo del análisis se debe conocer inicialmente cuales son los objetivos estratégicos organizacionales para identificar su direccionamiento y metas propuestas.

Actualmente existe un indicador clave para medir la efectividad de la recuperación de cartera el cual mide la calidad de cartera de exempleados:

Figura 5 Indicador calidad cartera total consumo + vivienda

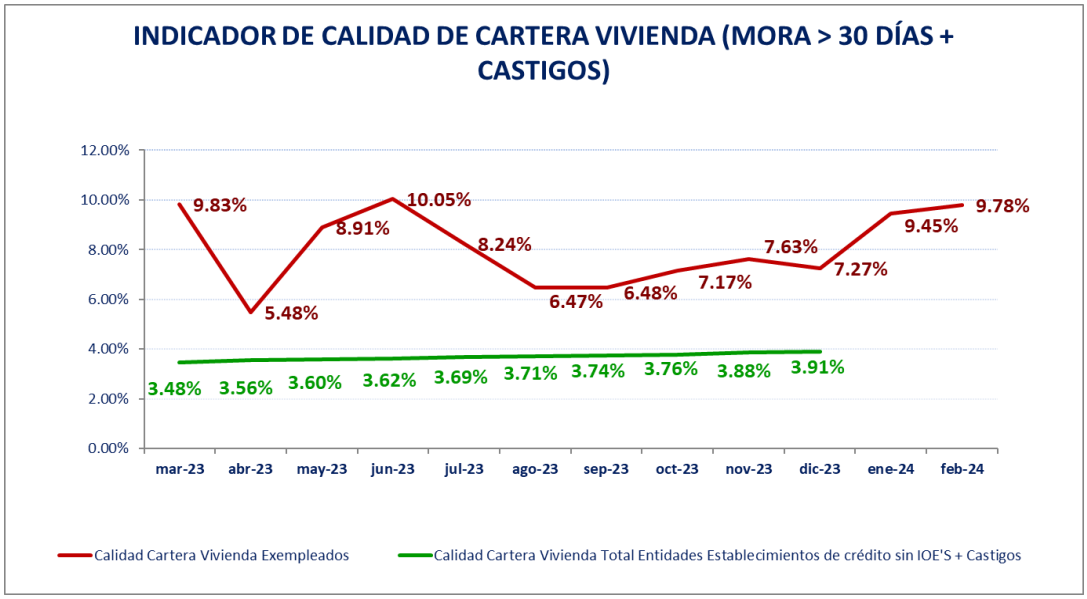


Nota. Indicador calidad cartera marzo 2023 a febrero 2024. Fuente: elaboración propia.

Como se observa en la gráfica anterior la calidad de cartera de exempleados es comparada con la calidad de cartera total en los establecimientos de crédito sin (Instituciones oficiales especiales), en las últimas doce mediciones de este indicador la EFoe superó a los

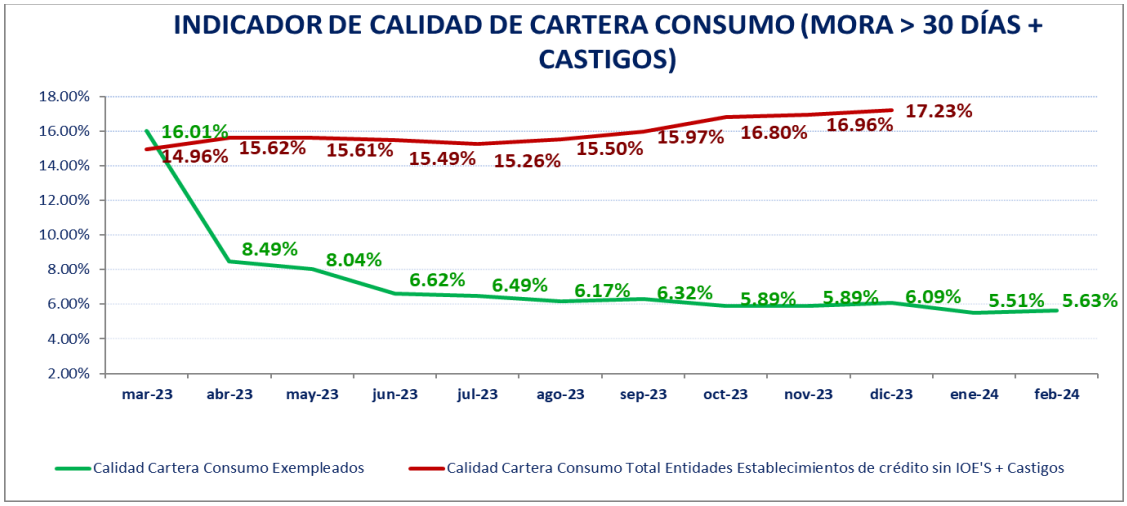
establecimientos de crédito en doce periodos alcanzando un indicador máximo de calidad de cartera de 10.06%.

Figura 6 Indicador calidad cartera vivienda + castigos



Nota. Indicador calidad cartera vivienda + castigos marzo 2023 a febrero 2024. Fuente: elaboración propia.

Figura 7 Indicador calidad cartera consumo + castigos



Nota. Indicador calidad cartera consumo + castigos marzo 2023 a febrero 2024.

Fuente: elaboración propia.

Otros indicadores importantes son la calidad de cartera de vivienda (mora >30 días + castigos) y calidad de cartera consumo (mora >30 días + castigos). Para vivienda se encuentra que durante todo el periodo analizado el indicador fue superior al total de entidades de crédito sin IOE´S + castigo y para consumo en una medición superó al total de entidades de crédito sin IOE´S + castigo.

8.1 Foda

Figura 8 Matriz Foda

		FORTALEZAS		DEBILIDADES
		CONCEPTO		CONCEPTO
ASPECTOS INTERNOS	F1	Alto conocimientos del deudor y habito de pago	D1	Tasas de intereses moratorias bajas para la línea de vivienda
	F2	Atención personalizada y acercamiento constante al deudor	D2	No hay cobros adicionales de gestión de cobranza (gastos procesales, honorarios de abogado)
	F3	Creacion y ejecución del plan de cobranza mensual	D3	Existe un único gestor de cobranza, con un perfil comportamental y relación previa con los deudores
	F4	Seguimiento constante a los compromisos de pago de los deudores	D4	Poco compromiso de la deuda
	F5	Canales de contacto efectivos para la gestión de cobranza	D5	No se otorgan condonaciones de intereses
	F6	Capacitación constante en temas de gestión y recupeación de cartera	D6	No existe un sistema de gestión de cobranza
	F7	Información actualizada y veraz sobre la situación del deudor	D7	Crecimiento mora mayor a 30 días
			D8	Pago irregular de las cuotas de un deudor
			D9	Concentración de gestión de cobranza
			D10	No atención a los canales de cobranza
			D11	Pago irregular de las cuotas de deudores que generan alto impacto
			D12	Poco compromiso de la deuda en altura de mora superior a 60 días
			D13	Información de los deudores desactualizada
		OPORTUNIDADES		AMENAZAS
		CONCEPTO		CONCEPTO
ASPECTOS EXTERNOS	O1	Los deudores desean ponerse al día en la obligación	A1	Desempleo o deterioro de la capacidad de pago
	O2	Nueva herramienta de cobro por SMS	A2	Poco conocimiento financiero de parte de los deudores
	O3	Profesionales de cartera capacitados para realizar gestión de cobranza	A3	Incremento en la tasa de usura
	O4	Contratación de sistema de gestión de cobranza en tramite	A4	Priorización de pago de otros créditos, por tener tasas más altas
			A5	Incremento días de mora
			A6	Deudores que fueron normalizados o mejoraron su rango de mora vuelvan a su situación antigua
			A7	El aumento del IPC 2023 y las alzas en la economía generen aumento en pago de créditos externos
			A8	Ley 2300 de 2023 pone limitaciones al proceso de recuperación de cartera

Nota. Matriz Foda proceso gestión y recuperación de cartera exempleados.

Fuente: elaboración propia.

8.2 Identificación del problema de negocio

Explicado lo anterior se plantean los diferentes problemas encontrados dentro del proyecto:

- A la fecha no se encuentra establecido una metodología de estudio de crédito que permita el análisis de diferentes variables y separe los deudores buenos de los malos para predecir el incumplimiento en el pago de los créditos de la cartera de exempleados de la EFoe.
- La gestión de recuperación de cartera se realiza a través de un plan de gestión de cobranza mensual, donde se gestionan a todos los deudores de la cartera de exempleados desde el día 1 de mora, se hacen segmentaciones y se gestionan por los diferentes canales de contacto autorizados (llamadas, mensajes de texto, correos electrónicos, whatsapp business), esta gestión se ha ajustado para dar cumplimiento a la ley 2300 de 2023 y es aquí donde un modelo predictivo de recuperación de créditos cobra importancia pues permitiría gestionar a los deudores que presente una alta probabilidad de incumplimiento, ahorrando tiempo, dinero en gestión y segmentando de una mejor manera a los deudores.
- Adicionalmente en la EFoe, las carteras de empleados y exempleados, han crecido de una manera muy rápida y comparándonos con otras entidades del sector, no encontramos ninguna entidad tenga una cartera tan amplia como beneficio a sus empleados, hemos encontrado que en bancos muy grandes como Bancolombia o Davivienda, la cartera máxima que se otorga a empleados es de 50,000 millones de pesos, número que ya fue superado ampliamente en la EFoe.
- El indicador de calidad de cartera vencida de la cartera de exempleados ha experimentado cambios significativos en su resultado, este indicador mide la cartera de vivienda, consumo y vivienda + consumo y es comparado con el promedio del total de los

establecimientos de crédito del país, durante el 2023 y lo corrido del 2024 el indicador de calidad de cartera total (consumo + vivienda) alcanzó un valor máximo de 10,05% mientras el resultado máximo de las entidades de crédito fue del 6,5%, superar este indicador es una alerta para la Efoe.

8.3 Problema analítico a resolver

La pregunta de analítica de datos que se intenta responder es cuáles son los deudores que de la cartera de exempleados incumpliran en el pago la cuota de su crédito en el siguiente mes. Esto con el fin de poder hacer una gestión de cobranza diferencial tomando decisiones sobre qué deudores se deben gestionar y cuáles no, de esta manera se ahorra tiempo y se puede realizar una gestión completa y más estratégica a los deudores con alta probabilidad de incumplimiento. Para dar solución a este problema es importante determinar los objetivos de minería del negocio:

- Analizar a profundidad el contexto y las necesidades específicas de La EFoe para identificar las necesidades y oportunidades claves para el otorgamiento de créditos a empleados y la gestión de recuperación de cartera de exempleados, alineando el proyecto con la estrategia y alcance del negocio.
- Realizar la recopilación, descripción y exploración de datos, con el fin de entender y permitir un análisis de los datos históricos de la cartera de empleados y exempleados de la EFoe.
- Diseñar un modelo predictivo de scoring de recuperación de créditos y de machine learning que permitan una evaluación precisa y en tiempo real de los riesgos de crédito, facilitando la toma de decisiones informadas y oportunas teniendo en cuenta el perfil de riesgo crediticio de La EFoe.

- Determinar la viabilidad de implementación del modelo de riesgo para la cartera de empleados y exempleados de La EFoe y proporcionar conclusiones o recomendaciones basadas en los resultados del modelo de riesgo.

8.4 Solución propuesta a la EFoe

Se propone a la EFoe comenzar con una identificación adecuada de los clientes, para ubicarlos en un grupo prioritario para la gestión de recuperación de cartera. De esta manera, la Dirección de Cartera podrá concentrar sus esfuerzos y recursos en este segmento. Esto se podrá lograr desarrollando un modelo predictivo de scoring y modelos de machine learning para la EFoe que permita anticipar el riesgo de impago de un crédito, de esta manera, el modelo servirá como base para el diseño de estrategias de contención, gestión y en determinados casos aceleración de las obligaciones crediticias, convirtiéndose así en una herramienta de gran utilidad para la gestión y recuperación de cartera de la EFoe. Entre las características principales del producto a entregar se encuentra la generación de valor para la Dirección de Cartera, a través de los resultados obtenidos, se fortalecerá el seguimiento continuo a los deudores de la cartera de exempleados. Además, esta herramienta será útil para la dirección al tomar decisiones de carácter funcional y estratégico.

9 Descripción de las alternativas, estrategias y/o acciones para dar solución a la problemática

9.1 Ética, seguridad y aspectos legales de los datos

Es importante abordar el capítulo referente a la ética, seguridad y aspectos legales de los datos, para el desarrollo de este proyecto se tomaron diferentes fuentes de información y variables relevantes para poder desarrollar un modelo predictivo de scoring para recuperación de créditos, estos datos deben ser clasificados y tratados de acuerdo a las diferentes leyes que rigen en Colombia sobre el tratamiento de los datos, entre las cuales se encuentran: ley 1266 de 2008 Habeas Data, ley 1581 de 2012 Ley de Protección de Datos Personales.

Primero es importante mencionar que todos los empleados y exempleados de la EFoe firmaron una autorización expresa en un documento de solicitud de crédito, para que sus datos fueran tratados de acuerdo a la ley 1581 de 2012, de esta manera los datos son almacenados y custodiados por la EFoe y a su vez la información relevante a sus crédito será compartida y actualizada mensualmente, la política menciona lo siguiente : “En mi calidad de Titular de la información, actuando libre y voluntariamente, autorizo a La EFoe y/o a terceros contratados por la EFoe, o quien represente sus derechos, a acceder a mis datos personales contenidos en la base de datos de Operadores de información de seguridad social autorizados por el Ministerio de Salud y Protección Social o de administradoras de pensiones, a mis datos personales recolectados por medio del presente formulario, y a mis datos personales contenidos en las bases de datos de los operadores de información crediticia, en adelante mi información personal, para darle tratamiento en los términos expresados en la Política de Tratamiento de la Información Personal y para finalidades de gestión de riesgo crediticio tales como: (i) elaboración y circulación a terceros de scores crediticios,

herramientas de validación de ingresos, herramientas predictivas de ingresos, herramientas para evitar el fraude y en general**,** herramientas que permitan adelantar una adecuada gestión del riesgo crediticio. (ii) Compararla, contrastarla y complementarla con la información financiera, comercial, crediticio, de servicios y proveniente de terceros países. (Formato público de la EFoe).

Lo anterior implica que el cumplimiento o incumplimiento de mis obligaciones se refleja en las mencionadas bases de datos, en donde se consignan de manera completa, todos los datos referentes a mi actual y pasado comportamiento frente al sector Financiero y en general, frente al incumplimiento de mis obligaciones.

Así mismo, en cumplimiento de la Ley de tratamiento de datos personales, Ley 1581 de 2012 y su Decreto Reglamentario 1377 de 2013 autorizo a la EFoe a administrar la información suministrada en la presente solicitud de préstamo".

Una vez se tiene esta autorización es importante clasificar los datos de acuerdo:

- Datos públicos: como aquellos contenidos en documentos públicos
- Datos privados: es decir, aquellos que "por su naturaleza íntima o reservada sólo son relevantes para el titular
- Datos semi privados: como aquellos que no tiene "naturaleza íntima, reservada, ni pública y cuyo conocimiento o divulgación puede interesar no sólo a su titular sino a cierto sector o grupo de personas o a la sociedad en general
- Datos sensibles: En el caso no se advierte que estén involucrados datos como los definidos en el artículo 5 de la ley 1581 de 2012 que pudiesen afectar la intimidad o generar algún tipo de discriminación. (Superintendencia de Industria y Comercio, 2023)

Tabla 1 Clasificación datos

Datos Públicos	Datos privados	Datos semiprivados	Datos sensibles
Número de identificación (cedula)		Número de crédito	
Nombres y apellidos		Valor crédito	
Género		Plazo crédito	
		Fecha desembolso	
		Plazo residual	
		Saldo Capital	
		Motivo salida	
		fecha de salida	
		Meses desde el retiro	
		Meses desde la aprobación del crédito	
		Fecha de nacimiento	
		Edad	
		Endeudamiento	
		Inicio empleado	
		Antigüedad cuando otorgaron el crédito	
		Tiempo trabajado	
		Cargo	
		Tasa	
		Valor capital	
		Valor intereses	
		Cuota total	
		Cuentas x Cobrar	
		Cuotas vencidas a la fecha de corte	
		Días Mora	
		Calificación de riesgo actual	
		Clase	
		Línea de Crédito	
		Portafolio Cobranza	
		Provisión General	
		Estado Crediticio	
		Portafolio	
		Mora últimos 6 meses	
		Mora últimos 3 meses	

Nota. Clasificación de datos utilizados en el proyecto empresarial.
Fuente: elaboración propia.

Como se muestra en la tabla anterior para el desarrollo del proyecto no se requieren datos privados ni sensibles, sin embargo, número de identificación, nombres y apellido, pueden hacer una fácil identificación de los demás datos de un determinado deudor, es por esto que no serán tenidos en cuenta, adicionalmente el sexo (hombre o mujer) al ser considerada una variable que puede discriminar y sesgar tampoco será incluida.

Adicionalmente y aunque los datos de nombres - apellidos y número de identificación no van a ser tenidos en cuenta en el modelo en caso de ser necesario se debe plantear una anonimizarían para estas dos variables:

Tabla 2 Anonimización

Nombres	**Nombre anonimizado**
Juana Alberta ="$&%"&IZQUIERDA(Nombres;2)	$&%Ju
Apellidos	**Apellido anonimizado**
Labra ="$&%"&IZQUIERDA(Apellido;2)	$&%La
Número de identificación	**Número de identificación**
5572332 ="*"&DERECHA(Número de identificación;3)**	***332

Nota. Anonimización propuesta. Fuente: elaboración propia.

9.2 Analytics life cycle management

9.2.1 Comprensión de los datos

9.2.1.1 Recolección de Datos Iniciales.

Los datos iniciales utilizados en este proyecto son las bases que tiene La EFoe para las carteras de empleados y exempleados desde el año 2000 hasta el 2023.

La información será extraída de cuatro fuentes principales las cuales deberán ser compiladas en una base de datos principal, el dato que servirá para unificar toda la información es el número de crédito este número es único y no se repite. A continuación, se describen las diferentes fuentes de datos que se utilizaran en el proyecto:

- **Informe de cartera:** Esta base de datos es extraída de un programa licenciado de la EFoe en el cual encontramos las siguientes variables: Nro Solicitud/Crédito, No Pagare, Fecha Expedición, Fecha Desembolso del crédito, Valor Inicial Cartera, Saldo Cartera ($), Tipo de tasa, Mod. Pag.(K), Mod. Pag.(I), Beneficiario, Fuente de Recursos, Plazo En Meses, Línea de Producto, Portafolio Cobranza

- **Informe de evaluación de cartera:** Esta base es extraída de un programa licenciado de La EFoe en el cual encontramos las siguientes variables: Crédito No. /Pagaré, Crédito No, Tipo de Identificación, No. Iden., Deudor/Beneficiario, Saldo Capital, Intereses Corrientes en Balance, Intereses Mora en Cuentas de Orden, Fecha Venc. Cuota, Días Mora, Días Después del Incumplimiento, Calif. Ope. Sup. Act., Clase, Provisión CIP Capital, Provisión CIP Intereses, Línea de Crédito, Tipo Garantía.

- **Informe de altura de mora histórico desde 2000:** Este informe es construido por la Dirección de Cartera de manera mensual y las variables que incluye son las siguientes: Fecha Corte, Crédito No. Pagaré, Tipo de Identificación, Número de Identificación, Deudor/Beneficiario, Saldo Capital, Días Mora, Línea de Crédito, Portafolio.

- **Repositorio de información documental de créditos:** Se tiene actualmente un programa de repositorio documental en el cual se encuentra la información laboral de los empleados como contrato (fecha de ingreso, cargo, nivel educativo, ingresos al momento de otorgamiento del crédito), información personal (fecha de nacimiento, ciudad de nacimiento, género, ciudad de nacimiento, estrato de la vivienda actual), familiar (estado civil, tipo de vivienda, personas a cargo), endeudamiento al solicitar el crédito. Adicionalmente para los exempleados información de fecha de retiro de la entidad y motivo de retiro.

9.2.1.2 Descripción de las Variables.

A continuación, se hace una descripción de todas las variables de datos iniciales:

- Número de crédito: dato de tipo numérico, representa el número único de la obligación de crédito inicia por 3 y debe contener 14 dígitos.
- Fecha desembolso: dato tipo fecha, es la fecha en la cual se aprobó el crédito, debe contener 8 dígitos y 2 caracteres especiales (/), es del tipo día/mes/año.
- Valor crédito: dato de tipo numérico, indica el valor total que fue desembolsado, valor mínimo 1,000,000 máximo 1,500,000,000.
- Saldo Capital ($), saldo capital: dato de tipo numérico, indica el valor total del saldo del crédito valor mínimo 1,000,000 máximo 1,500,000,000.
- Plazo crédito: dato de tipo numérico, indica el número de meses en los que se va a cancelar el crédito, los valores deben ser 48, 84 0 240.
- Plazo residual: dato de tipo numérico, indica el número de meses que faltan para cancelar el crédito a una determinada fecha de corte, los valores deben estar entre 1 y 240.

- Línea de crédito: dato de tipo alfanumérico, indica si el crédito es de vivienda o consumo.
- Portafolio Cobranza: dato de tipo alfanumérico, indica en qué portafolio de cobranza se encuentra el crédito en un cierre determinado, los valores pueden ser, ordinaria, persuasiva, perjudicial, jurídica.
- Días Mora: dato de tipo numérico, indica el número de días después que se venció la fecha de pago de una cuota determinada, dato no debe contener más de 4 dígitos.
- Calificación de riesgo actual: dato de tipo alfanumérico, la calificación actual de riesgo de crédito, los valores pueden ser A, B, C, D o E.
- Cuota total: es el valor de capital e intereses que debe pagar un deudor en determinada fecha de corte, es un dato de tipo numérico.
- Cuentas por cobrar: dato de tipo numérico, indica el valor pendiente por cobrar de concepto seguros de vida deudor.
- Provisión general: dato de tipo numérico, indica el valor de provisión que genera el crédito.
- Cuotas vencidas a la fecha de corte: variable numérica, indica la cantidad de cuotas vencidas que tenía el deudor en determinada fecha de corte.
- Estado crediticio: variable de tipo alfanumérico, puede tener los siguientes valores: vigente, modificado o reestructurado.
- Clase: dato de tipo alfanumérico, indica si el crédito es de vivienda, libre inversión, vehículo, estudio o calamidad doméstica.

- Fecha Corte: dato tipo fecha, es la fecha en la cual se producen diferentes reportes de información en un corte determinado, debe contener 8 dígitos y 2 caracteres especiales (/), es del tipo día/mes/año.
- Fecha de inicio empleado: dato tipo fecha, es la fecha en la cual el empleado entró a planta en La EFoe, debe contener 8 dígitos y 2 caracteres especiales (/), es del tipo día/mes/año.
- Cargo: dato de tipo alfanumérico, indica el tipo de cargo del empleado los valores posibles son secretaria, analista, profesional, gerente, directivo, vicepresidente o presidente.
- Fecha de nacimiento: dato tipo fecha, es la fecha de nacimiento del deudor, debe contener 8 dígitos y 2 caracteres especiales (/), es del tipo día/mes/año.
- Endeudamiento: dato de tipo numérico, indica el dinero que paga el empleado en obligaciones financieras, el dato no debe contener más de 5 dígitos y está expresado en miles es decir 1240 indica 1,240,000 de endeudamiento, (se actualiza cada año).
- Fecha de salida de la entidad: dato tipo fecha, es la fecha de retiro de un empleado, el momento en que se convierte en exempleado, debe contener 8 dígitos y 2 caracteres especiales (/), es del tipo día/mes/año.
- Motivo de retiro: dato de tipo alfanumérico, indica el motivo de retiro de la entidad, los valores pueden ser retiro voluntario o despido.

9.2.1.3 Análisis Descriptivo de los Datos.

El proceso de comprensión y preparación de los datos del presente Proyecto Empresarial inició con la clasificación de las variables en dos grupos.

Por una parte, se agruparon todas las variables numéricas, con el objetivo de hacer un análisis con los principales indicadores estadísticos (Valores mínimos, valores máximos, media, desviación estándar, varianza y curtosis); con el objetivo de verificar la calidad de los datos de estas variables.

A continuación, se relaciona el cuadro de Estadísticos descriptivos de este primer grupo.

Figura 9 Estadísticos descriptivos iniciales

	Valor crédito	Plazo crédito	Plazo residual	Saldo Capital	Edad	Endeudameinto	Antigüedad cuando otorgaron el crédito	Tiempo trabajado
count	145	145	145	145	145	145	145	145
mean	293,797,000	205	112	179,524,500	49	3,590,600.00	6	8
min	20,000,000	48	5	2,731,339	32	-	0	1
25%	125,100,500	240	74	53,287,500	39	780,000.00	2	4
50%	265,000,000	240	109	122,984,500	48	1,791,000.00	3	5
75%	400,000,000	240	165	242,426,500	58	3,907,000.00	7	10
max	1,030,000,000	240	233	947,943,500	73	65,177,000.00	32	32
std	211066400.00	70.06	58.93	182693300.00	10.90	6746048.00	6.83	7.16

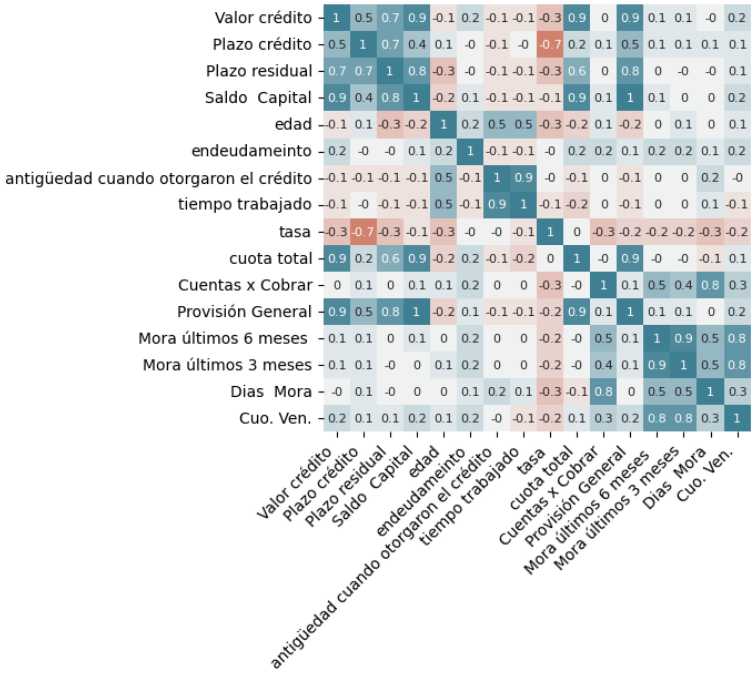
	Tasa	Cuota total	Cuentas x Cobrar	Cuotas vencidas	Dias Mora	Provisión General	Mora últimos 6 meses	Mora últimos 3 meses
count	145	145	145	145	145	145	145	145
mean	0.021	1,657,646	221,252.20	0	55	1,718,178.00	1	0
min	0.000	128,665	-	0	0	-	0	0
25%	0.016	951,983	7,323.00	0	0	367,286.80	0	0
50%	0.016	1,384,166	18,940.00	0	0	1,178,374.00	0	0
75%	0.016	2,101,224	91,737.00	0	0	2,424,264.00	0	0
max	0.065	7,893,852	7,372,658.00	3	3889	9,479,435.00	6	3
std	0.01	1126000.00	801807.80	0.53	400.16	1886217.00	1.43	0.81

Nota. Estadísticos base de datos inicial, elaborado en Python. Fuente: elaboración propia.

- Cantidad mínima de registros: Se observa que las variables que se relacionan tienen una cantidad suficiente de datos.

- Valores mínimos y máximos: Se realiza la revisión de estos valores con respecto a la definición de cada variable y se concluye que no existen valores incoherentes y que la información registrada para cada variable es precisa.

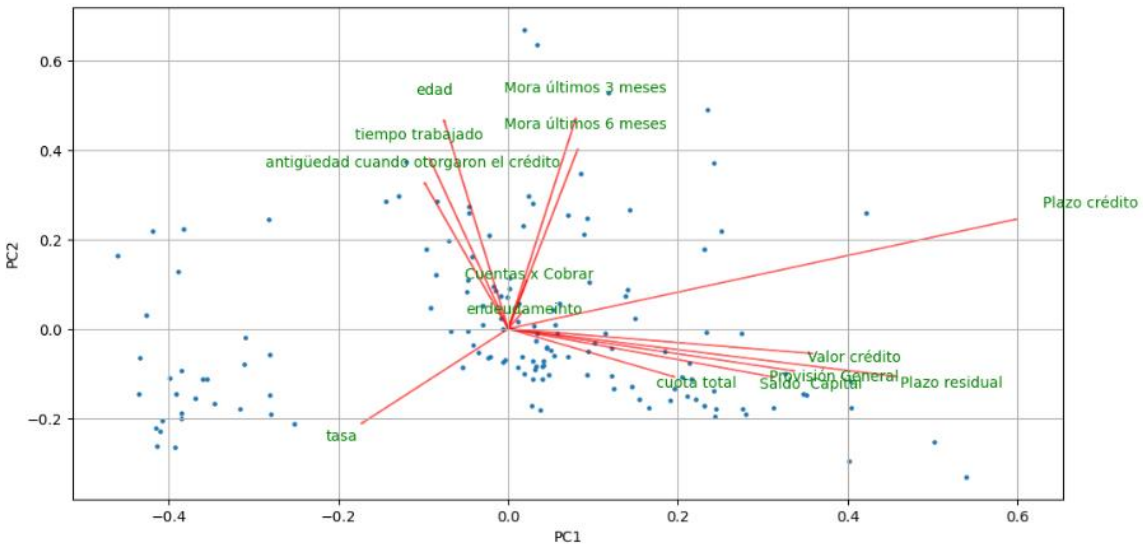
Figura 10 Matriz de correlación de variables



Nota. Matriz de análisis de correlación de variables, elaborado en Python.

Fuente: elaboración propia.

Figura 11 Análisis de componentes principales



Nota. Ilustración componentes principales base de datos, elaborado en Python.

Fuente: elaboración propia.

9.2.1.4 Revisión de los Datos.

Con base en la sección anterior es importante validar aspectos clave como si existen datos nulos o vacíos, si existen datos duplicados, si existen datos atípicos o incongruentes y que tratamiento se les dio. A continuación, se relacionan las diferentes variables y se explica que limpieza de datos se le realizó:

- Primero se validó que no existieran datos duplicados y el resultado fue conforme, no se visualizan datos de número de crédito repetidos, este valor es único para cada crédito.
- No se encuentran dos datos atípicos
- No se encuentran datos vacíos
- Se encuentran 5 registros de deudores con variables vacías por lo cual se eliminan estos deudores pues no es posible calcular o determinar los valores de las variables vacías.

9.2.2 Preparación de los datos

Como se mencionó anteriormente esta fase debe seleccionar los datos, realizar su respectiva limpieza, realizar ingeniería de características, dar formato a los datos e integrarlos.

9.2.2.1 Construcción de Variables.

Usando la información obtenida se construyeron las siguientes variables:

- Edad: Se obtiene la edad de los empleados en el momento de desembolso del crédito a partir de la fecha de nacimiento y del día de desembolso del crédito
- Endeudamiento se convierte en miles se multiplica cada dato por 1000

- Meses desde el retiro de la empresa: se calcula utilizando la fecha de salida de la empresa y la fecha de corte así se conoce cuantos años han pasado desde el retiro a hoy (fecha de corte)
- Meses desde la aprobación del crédito: se calcula utilizando la fecha de desembolso del crédito y la fecha de corte así se conoce cuantos años han pasado desde el desembolso a hoy (fecha de corte)
- Antigüedad cuando le otorgaron el crédito: se calcula utilizando la fecha de desembolso del crédito y la fecha de inicio empleado así se conoce cuantos años de antigüedad en la empresa tenía el empleado cuando le otorgaron el crédito.
- Tiempo trabajado: se calcula utilizando la fecha de retiro de la empresa y la fecha de inicio empleado así se conoce cuantos años de antigüedad en la empresa tenía el empleado cuando se retiró.
- El valor de la cuota mensual se obtiene de dividir el valor del monto total del crédito en el número de meses del plazo, luego se multiplica por la tasa de intereses y se obtiene el valor mensual de la cuota.
- Mora últimos tres meses: variable numérica, se construye revisando los últimos tres cortes 1 indica que en los últimos tres meses cerró en mora 1 vez, 2 que en los últimos tres meses cerró en mora 2 veces y 3 que en los últimos tres meses cerró 3 veces en mora.
- Mora últimos seis meses: variable numérica, se construye revisando los últimos seis cortes 1 indica que en los últimos seis meses cerró en mora 1 vez, 2 que en los últimos seis meses cerró en mora 2 veces, 3 que en los últimos seis meses cerro 3 veces en mora y así sucesivamente, valor máximo seis que indica en los últimos seis meses cerra 6 veces en mora.

- Variable objetivo: la variable que busca predecir o explicar el modelo es riesgo de crédito y arroja dos resultados 1 o 0; 0 indica que el deudor no presentará mora en el pago de su próxima cuota y 1 indica que si lo hará.

9.2.2.2 Seleccionar los Datos.

Una vez ubicadas todas las variables y datos se seleccionan los siguientes:

- Crédito
- Valor crédito
- Plazo de crédito
- Plazo residual
- Saldo de capital
- Motivo salida
- Meses desde el retiro
- Meses desde la aprobación del crédito
- Edad
- Endeudamiento
- Antigüedad cuando otorgaron el crédito
- Tiempo trabajado
- Cargo
- Tasa
- Valor capital
- Valor intereses
- Cuota total
- Cuentas por cobrar

- Cuotas vencidas
- Días mora
- Calificación de riesgo actual
- Clase
- Línea de crédito
- Provisión general
- Estado del crédito
- Mora últimos tres meses
- Mora últimos seis meses

9.2.2.3 Calidad de Datos.

Al revisar la calidad de los datos y analizar la descripción de los datos se puede interpretar que los datos son congruentes no se encuentran valores atípicos o que no guarden sentido.

En cuanto a la calidad de los datos la base fue construida de manera automática para las variables número de crédito, valor crédito, plazo de crédito, saldo de capital, tasa, cuentas por cobrar, cuotas vencidas, días mora, clase, línea de crédito, provisión general, estado del crédito y calificación de riesgo actual; por otro lado, fue necesario construir variables de manera manual, pero se aseguró que la construcción de las variables: edad, plazo residual, meses desde el retiro, meses desde la aprobación del crédito, antigüedad cuando otorgaron el crédito, tiempo trabajado, valor capital, valor intereses, cuota total, mora últimos tres meses, mora últimos seis meses tengan calidad y veracidad en la información. Por último, se trasladaron las siguientes variables para cada deudor: motivo salida, endeudamiento y cargo.

9.2.3 Modelado.

En esta fase de la metodología se escogerá la técnica (o técnicas) más apropiadas para los objetivos marcados de la minería de datos. A continuación, y una vez realizado un plan de prueba para los modelos escogidos, se procederá a aplicar dichas técnicas sobre los datos para generar el modelo y por último se tendrá que evaluar si dicho modelo ha cumplido los criterios de éxito o no.

9.2.3.1 Escoger la Técnica de Modelado.

Se realizará un modelo de scoring de recuperación de créditos a partir de la metodología de regresión logística.

9.2.3.2 Metodología y validación del modelo.

Para realizar un modelo score es necesario determinar cuáles son deudores buenos y cuales son malos para esto, se clasificaron como buenos los que no han tenido mora mayor o igual a 30 días en el corte inmediatamente anterior a la fecha de corrida del cierre y serán clasificados como malos los que sí presentaron mora mayor o igual a 30 días en el corte inmediatamente anterior a la fecha de corrida del cierre.

$$Tasa\ de\ mora_i = \left(\frac{Malos_i}{Malos_i + Buenos_i}\right)$$

9.2.3.3 Análisis Univariante.

- Se han agrupado cada una de las variables en buckets y se han obtenido los WOES de manera que se maximice el Information Value.

9.2.3.4 Cargue y procesamiento de la base de datos.

Después de cargar la base en R Studio es necesario convertir ciertas variables cualitativas en factores que permitan el análisis estadístico, las variables que se convierten en factor fueron: motivo de salida, cargo, calificación de riesgo, línea de crédito.

Para la construcción de este modelo fue necesario determinar cómo variable dependiente, el incumplimiento en los pagos del crédito adquirido en la EFoe (créditos con mora posterior a los 30 días) esta variable es denominada riesgo de default, se encuentra que del total de la base 130 deudores no presentaron mora (0) y 15 si lo hicieron (1). Al tener este resultado es necesario realizar una técnica de sobremuestreo (oversampling) de la clase minoritaria (clase 0), se iguala el 15 a 130, el código realiza sobremuestreo de las observaciones en el conjunto de datos original para equilibrar las clases basadas en la variable de respuesta. Luego, crea una copia del conjunto de datos sobre muestreado y muestra la distribución de clases después del sobremuestreo. los datos balanceados hacen que se pueda trabajar correctamente un modelo.

Continuando con el desarrollo del modelo se testea con un 20% y se entrena con el 80% de los datos, esta división es comúnmente utilizada en el aprendizaje supervisado para entrenar un modelo en un conjunto de datos y evaluar su rendimiento en un conjunto independiente.

Tabla 3 Análisis WOE - IV Variables modelo

• Valor Crédito:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
Valor Crédito	[-Inf,140000000)	42	10	32	1,163	0.363
	[140000000,250000000)	42	22	20	-0.09531	0.363
	[250000000,350000000)	43	21	22	0.04652	0.363
	[350000000,544000000)	42	24	18	-0.2877	0.363
	[544000000, Inf)	43	29	14	-0.7282	0.363
• Plazo Crédito:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
Plazo Crédito	[-Inf,240)	35	10	25	0.9163	0.1537
	[240, Inf)	177	96	81	-0.1699	0.1537
• Plazo Residual:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
Plazo Residual	[-Inf,75.26666667)	42	10	32	1,163	0.4072
	[75.26666667,98.13333333)	36	21	15	-0.3365	0.4072
	[98.13333333,118.93333333)	46	24	22	-0.08701	0.4072
	[118.9333333,171.6)	41	29	12	-0.8824	0.4072
	[171.6, Inf)	47	22	25	0.1278	0.4072
• Saldo Capital:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
Saldo Capital	[-Inf,57594916)	42	10	32	1,163	0.4933
	[57594916,119822303)	39	15	24	0.47	0.4933
	[119822303,209644551)	46	24	22	-0.08701	0.4933
	[209644551,345432020)	39	28	11	-0.9343	0.4933
	[345432020, Inf)	46	29	17	-0.5341	0.4933

• Motivo Salida:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
Motivo Salida	[-Inf,2)	138	85	53	-0.4724	0.4237
	[2,3)	23	7	16	0.8267	0.4237
	[3, Inf)	51	14	37	0.9719	0.4237
• edad:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
edad	[-Inf,41.11780822)	40	8	32	1,386	0.6653
	[41.11780822,48.4630137)	44	19	25	0.2744	0.6653
	[48.4630137,53.46849315)	32	20	12	-0.5108	0.6653
	[53.46849315,59.50684932)	53	40	13	-1,124	0.6653
	[59.50684932, Inf)	43	19	24	0.2336	0.6653
• endeudamiento:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
endeudamiento	[-Inf,941000)	42	8	34	1,447	0.89
	[941000,1670000)	34	13	21	0.4796	0.89
	[1670000,3860000)	50	25	25	0	0.89
	[3860000,7682000)	41	22	19	-0.1466	0.89
	[7682000, Inf)	45	38	7	-1,692	0.89
• antigüedad cuando otorgaron _el_ crédito:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
antigüedad cuando otorgaron _el_ crédito	[-Inf,1.369444444)	42	29	13	-0.8023	0.5218
	[1.369444444,2.594444444)	38	10	28	1.03	0.5218
	[2.594444444,3.544444444)	47	29	18	-0.4769	0.5218
	[3.544444444,10.04722222)	39	11	28	0.9343	0.5218
	[10.04722222, Inf)	46	27	19	-0.3514	0.5218

• tiempo trabajado:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
tiempo trabajado	[-Inf,2.819444444)	42	30	12	-0.9163	0.3208
	[2.819444444,4.286111111)	41	15	26	0.55	0.3208
	[4.286111111,7.419444444)	44	16	28	0.5596	0.3208
	[7.419444444,12.08888889)	41	25	16	-0.4463	0.3208
	[12.08888889, Inf)	44	20	24	0.1823	0.3208
• cargo:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
cargo	[-Inf,3)	17	7	10	0.3567	0.011
	[3,9)	49	25	24	-0.04082	0.011
	[9, Inf)	146	74	72	-0.0274	0.011
• tasa:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
tasa	[-Inf, Inf)	212	106	106	0	0
• Cuentas _x_ Cobrar:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
Cuentas _x_ Cobrar	[-Inf,8329)	42	10	32	1,163	0.7972
	[8329,21077)	42	15	27	0.5878	0.7972
	[21077,53446)	30	10	20	0.6931	0.7972
	[53446,281414)	55	42	13	-1,173	0.7972
	[281414, Inf)	43	29	14	-0.7282	0.7972
• Calif._Ope._Sup._Act.:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
Calif._Ope._Sup._Act.	[-Inf,2)	163	61	102	0.5141	1,135
	[2, Inf)	49	45	4	-2.42	1,135
• Línea de Crédito:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
Línea de Crédito	[-Inf,3)	35	10	25	0.9163	0.1537
	[3, Inf)	177	96	81	-0.1699	0.1537

• Provisión General:						
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**
Provisión General	[-Inf,436297.67)	42	10	32	1,163	0.5157
	[436297.67,1198223.03)	41	15	26	0.55	0.5157
	[1198223.03,2096445.51)	44	24	20	-0.1823	0.5157
	[2096445.51,3454320.2)	39	28	11	-0.9343	0.5157
	[3454320.2, Inf)	46	29	17	-0.5341	0.5157

Nota. Tabla análisis univariante para determinar WOE e IV de las variables.

Fuente: elaboración propia.

Figura 12 Selección de variables por IV

variable	IV
tasa	0
cargo	0.011
Plazo Crédito	0.1537
Línea de Crédito	0.1537
tiempo trabajado	0.3208
cuota total	0.3597
Valor Crédito	0.363
Plazo Residual	0.4072
Motivo Salida	0.4237
Saldo Capital	0.4933
Provisión General	0.5157
antigüedad cuando otorgaron _el_ crédito	0.5218
edad	0.6653
Cuentas _x_ Cobrar	0.7972
endeudamiento	0.89
Calif._Ope._Sup._Act.	1,135

Se descartan las variables con IV menor a 0.02

variable	IV
tasa	0
cargo	0.011

Nota. Selección de variables significativas del modelo por IV.
Fuente: elaboración propia.

Posteriormente se realiza la transformación de los datos a partir del cálculo de los WOES.

Se realiza la regresión logística donde riesgo de default es la variable de respuesta y plazo crédito, plazo residual, endeudamiento, antigüedad cuando le otorgaron el crédito, tiempo trabajado, cuentas por cobrar y calificación de riesgo, son las variables predictoras, son significativas dentro del modelo y lo explican bien.

Figura 13 Selección de variables por IV

Significancia de las variables WOES y Regresión Logística	
Plazo Crédito Woe	*
Plazo Residual Woe	*
endeudamiento woe	*
antigüedad cuando otorgaron el crédito woe	***
tiempo trabajado woe	***
Cuentas_x_Cobrar_woe	**
Calif._Ope._Sup._Act._woe	**

Nota. Selección de variables significativas del modelo por regresión logística R Studio.
Fuente: elaboración propia.

Se comprueba la importancia de las variables y a partir de esto se puede concluir que no existe multicolinealidad entre las variables y que una variable no puede predecirse en gran medida a partir de la otra(s).

Como el modelo ya se estableció este se debe pasar a los datos de prueba y se arrojan los siguientes resultados:

Figura 14 Matriz de confusión

	test predi	
	0	1
0	22	3
1	8	17

Nota. Matriz de confusión. Fuente: elaboración propia

9.2.3.5 Estadísticas del Modelo de Scoring

Figura 15 Selección de variables por IV

Accuracy :	0.78
95% CI :	(0.6404, 0.8847)
No Information Rate :	0.6
P-Value [Acc > NIR] :	0.005688
Kappa :	0.56
Mcnemar's Test P-Value :	0.227800
Sensitivity :	0.7333
Specificity :	0.8500
Pos Prep Value :	0.8800
Neg Pred Value :	0.6800
Prevalence :	0.6000
Detection Rate :	0.4400
Detection Prevalence :	0.5000
Balanced Accuracy :	0.7917

Nota. Estadísticas del modelo por regresión logística R Studio.
Fuente: elaboración propia.

Una curva ROC (Receiver Operating Characteristic) del 78% generalmente se refiere a la AUC-ROC (Area Under the Curve of the Receiver Operating Characteristic) de un modelo de clasificación binaria. La AUC-ROC es una medida que evalúa la capacidad de un modelo para distinguir entre dos clases, y se representa gráficamente mediante la curva ROC.

La curva ROC es un gráfico que muestra la tasa de verdaderos positivos (sensibilidad) en el eje y y la tasa de falsos positivos (1 - especificidad) en el eje x. La AUC-ROC es el área bajo esta curva y proporciona una medida numérica de la capacidad de discriminación del modelo.

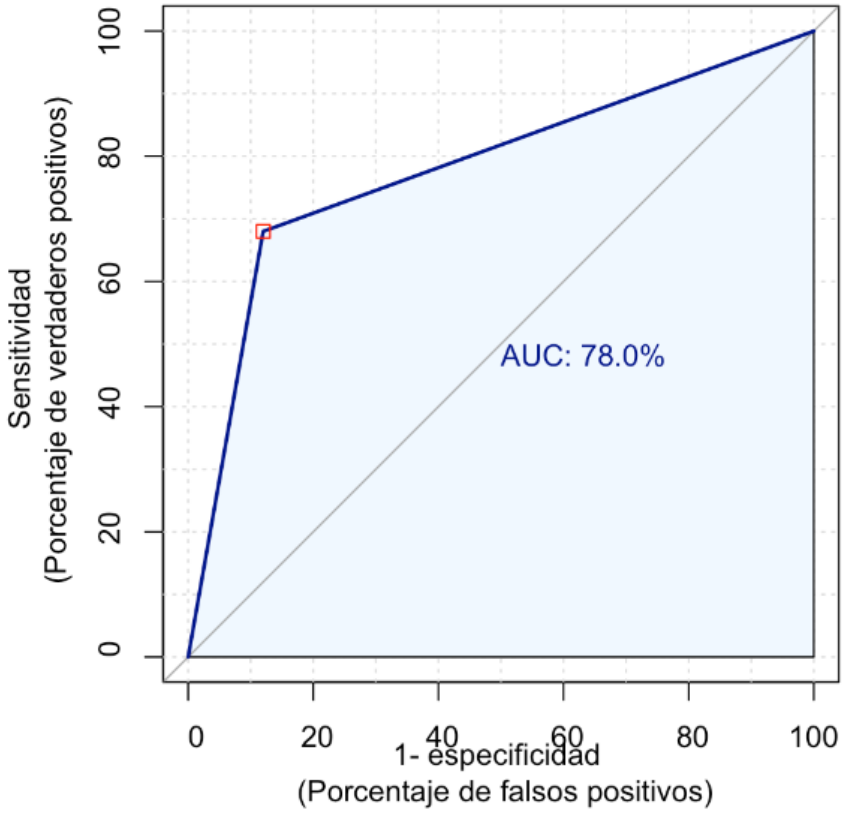
Un valor del 78% para la AUC-ROC indica que el modelo tiene una capacidad moderada para discriminar entre las dos clases. La interpretación de la AUC-ROC es la siguiente:

0.5: No hay capacidad de discriminación (como si el modelo estuviera prediciendo al azar).

0.7-0.8: Capacidad de discriminación buena.

Una AUC-ROC del 78% sugiere que el modelo tiene una capacidad razonable y ajustada para distinguir entre las clases.

Figura 16 Curva ROC Modelo scoring



Nota. Curva ROC modelo de scocing de riesgo de crédito R Studio.
Fuente: elaboración propia.

El siguiente paso es la creación del Balanced Scorecard de puntuación del modelo este permitirá clasificar a los deudores como buenos y malos (buenos indica que pagarán su próxima cuota de crédito y malos que no lo harán), utilizando las variables que resultaron significativas en el modelo.

Figura 17 Scorecard riesgo de no pago crédito

variable	bin	count	neg	pos	woe	total iv	points
Plazo Crédito	[-Inf,240)	35	10	25	0.9163	0.1537	-94
Plazo Crédito	[240, Inf)	177	96	81	-0.1699	0.1537	17
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**	**points**
Plazo Residual	[-Inf,75.26666667)	42	10	32	1,163	0.4072	67
Plazo Residual	[75.26666667,98.13333333)	36	21	15	-0.3365	0.4072	-19
Plazo Residual	[98.13333333,118.9333333)	46	24	22	-0.08701	0.4072	-5
Plazo Residual	[118.9333333,171.6)	41	29	12	-0.8824	0.4072	-51
Plazo Residual	[171.6, Inf)	47	22	25	0.1278	0.4072	7
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**	**points**
endeudamiento	[-Inf,941000)	42	8	34	1,447	0.89	61
endeudamiento	[941000,1670000)	34	13	21	0.4796	0.89	20
endeudamiento	[1670000,3860000)	50	25	25	0	0.89	0
endeudamiento	[3860000,7682000)	41	22	19	-0.1466	0.89	-6
endeudamiento	[7682000, Inf)	45	38	7	-1,692	0.89	-72
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**	**points**
antigüedad cuando otorgaron _el_ crédito	[-Inf,1.369444444)	42	29	13	-0.8023	0.5218	-64
antigüedad cuando otorgaron _el_ crédito	[1.369444444,2.594444444)	38	10	28	1.03	0.5218	83
antigüedad cuando otorgaron _el_ crédito	[2.594444444,3.544444444)	47	29	18	-0.4769	0.5218	-38
antigüedad cuando otorgaron _el_ crédito	[3.544444444,10.04722222)	39	11	28	0.9343	0.5218	75
antigüedad cuando otorgaron _el_ crédito	[10.04722222, Inf)	46	27	19	-0.3514	0.5218	-28

variable	bin	count	neg	pos	woe	total iv	points
tiempo trabajado	[-Inf,2.819444444)	42	30	12	-0.9163	0.3208	-84
tiempo trabajado	[2.819444444,4.286111111)	41	15	26	0.55	0.3208	50
tiempo trabajado	[4.286111111,7.419444444)	44	16	28	0.5596	0.3208	51
tiempo trabajado	[7.419444444,12.08888889)	41	25	16	-0.4463	0.3208	-41
tiempo trabajado	[12.08888889, Inf)	44	20	24	0.1823	0.3208	17
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**	**points**
Cuentas _x_ Cobrar	[-Inf,8329)	42	10	32	1,163	0.7972	58
Cuentas _x_ Cobrar	[8329,21077)	42	15	27	0.5878	0.7972	29
Cuentas _x_ Cobrar	[21077,53446)	30	10	20	0.6931	0.7972	35
Cuentas _x_ Cobrar	[53446,281414)	55	42	13	-1,173	0.7972	-59
Cuentas _x_ Cobrar	[281414, Inf)	43	29	14	-0.7282	0.7972	-36
variable	**bin**	**count**	**neg**	**pos**	**woe**	**total iv**	**points**
Calif._Ope._Sup._Act.	[-Inf,2)	163	61	102	0.5141	1,135	28
Calif._Ope._Sup._Act.	[2, Inf)	49	45	4	-2.42	1,135	-133

Se estima que el deudor con la mejor calificación de la cartera de empleados será el que tenga 365 puntos.

variable	points
base points	365

Nota. Balance Score card. Fuente: elaboración propia

9.2.4 Modelos de machine learning.

9.2.5 Balanceo de datos usando la técnica de Smote

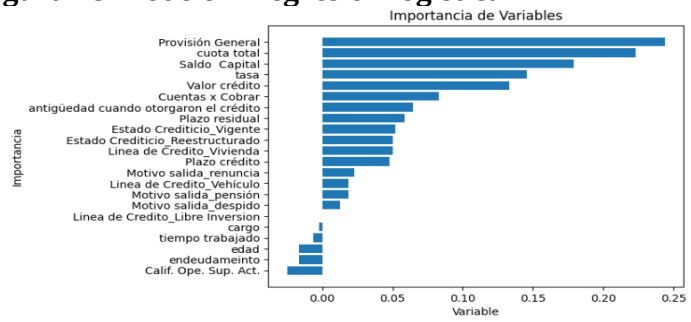
Una vez cargada la base en phyton se decide utilizar la técnica de balanceo smote, esta técnica es utilizada principalmente en datos donde una clase (deudores es mora) es significativamente menos frecuente que la otra (deudores sin mora), el smote aborda este problema generando nuevos ejemplos sintéticos de la clase minoritaria, en lugar de duplicar datos existentes o eliminar datos de la clase mayoritaria.

Para este ejercicio en particular el balanceo amplio la clase minoritaria a un 50%. Esto aumenta artificialmente el tamaño de la clase minoritaria y equilibra mejor las proporciones entre las clases, mejorando así el rendimiento del modelo de aprendizaje automático al aprender de manera más efectiva las características de ambas clases.

9.2.6 Selección de variables significativas para cada modelo

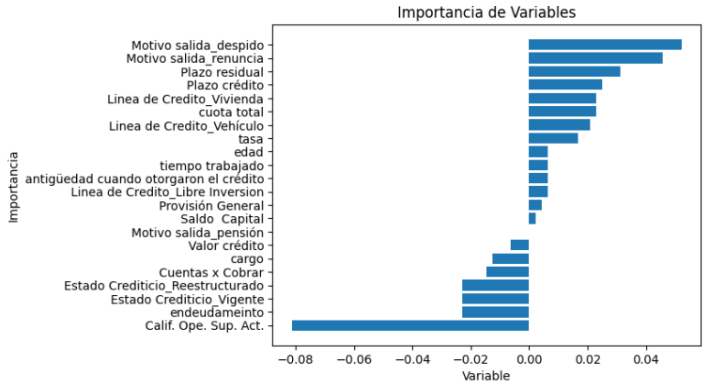
Una vez se corren los cinco modelos: Regresión Logística, Naive Bayes, Maquina de Soporte Vectorial, Árbol de clasificación y Random Forest, se evidencia que para cada uno las variables más significativas son diferentes y se evidencian en la siguiente gráfica:

Figura 18 Modelo I Regresión logística



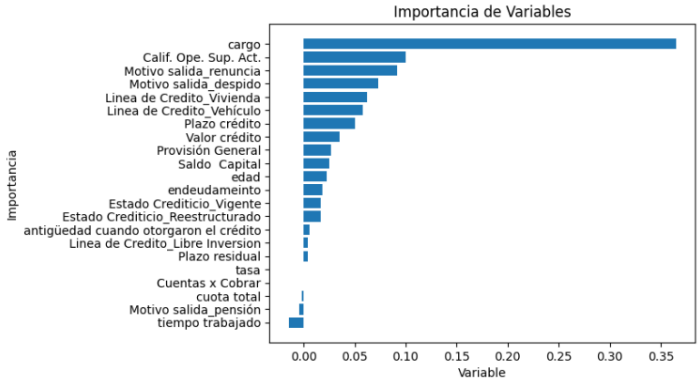
Nota. Variables significativas modelo I regresión logística Phyton.
Fuente: elaboración propia.

Figura 19 Modelo II Naibe Bayes



Nota. Variables significativas modelo II Naibe Bayes Phyton.
Fuente: elaboración propia.

Figura 20 Modelo III Máquina de soporte vectorial

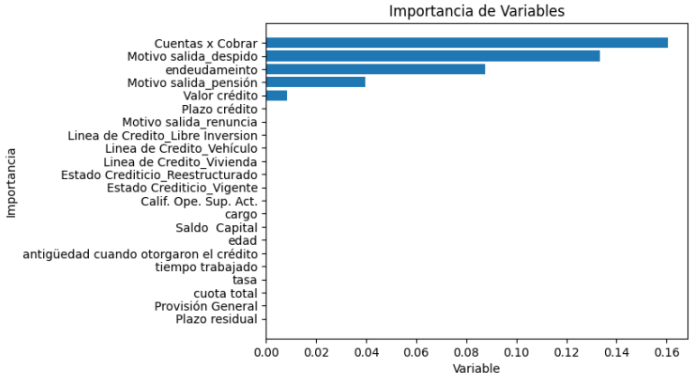


Nota. Variables significativas modelo I regresión logística Phyton.
Fuente: elaboración propia.

Figura 21 Modelo IV Random Forest

Nota. Variables significativas modelo IV Random Forest, Phyton
Fuente: elaboración propia.

Figura 22 Modelo IV Arbol de decisión



Nota. Variables significativas modelo IV Random Forest, Phyton
Fuente: elaboración propia.

9.2.7 Evaluación de los modelos

Para la selección del modelo se decide utilizar el recall como métrica más significativa porque es una métrica crucial cuando la identificación correcta de todos los casos positivos es de suma importancia, especialmente cuando los costes de los falsos negativos son elevados o cuando la clase positiva es minoritaria. Adicionalmente se tendrán en cuenta métricas como acurrancy, simplicidad, facilidad de adaptación a la EFoe, interpretación y mantenimiento

Figura 23 Evaluación de modelos

Modelo	Recall	Puntuación Recall	Accuracy	Puntuación Accuracy	ROC	Puntuación ROC	Simplicidad	Simplicidad
Random Forest	**0.97**	**6**	0.97	**6**	0.98	**6**	Bajo	**1**
Modelo de Scoring de Crédito	**0.73**	**2**	0.78	**3**	0.78	**3**	Moderado	**3**
Árbol de Clasificación	**0.85**	**4**	0.83	**4**	0.88	**4**	Moderado	**3**
Modelo de Regresión Logística	**0.75**	**3**	0.73	**1**	0.76	**2**	Alto	**5**
Modelo de Naive Bayes	**0.71**	**1**	0.74	**2**	0.68	**1**	Alto	**5**
Máquina de Soporte Vectorial (SVM)	**0.91**	**5**	0.91	**5**	0.9	**5**	Bajo	**1**

Modelo	Adaptabilidad	Adaptabilidad	Interpretación	Interpretación	Mantenimiento	Mantenimiento	Total
Árbol de Clasificación	Alto	**1**	Alto	**5**	Moderado	**3**	**28**
Modelo de Scoring de Crédito	Bajo	**5**	Alto	**5**	Bajo	**5**	**26**
Random Forest	Moderado	**3**	Moderado	**3**	Bajo	**5**	**26**
Modelo de Regresión Logística	Moderado	**3**	Alto	**5**	Bajo	**5**	**24**
Modelo de Naive Bayes	Moderado	**3**	Moderado	**3**	Bajo	**5**	**20**
Máquina de Soporte Vectorial (SVM)	Alto	**1**	Bajo	**1**	Alto	**1**	**19**

Nota. Evaluación de modelos con métricas, Phyton. Fuente: elaboración propia.

Según la gráfica anterior se puede evidenciar que existen dos modelos que por sus puntuaciones de métricas pueden considerarse como la mejor opción para implementar en la EFoe el primero es un modelo de machine learning denominado Árbol de clasificación y el segundo un modelo de scoring de crédito.

Debido a la madurez de la EFoe y su enfoque en la facilidad de explicabilidad, adaptabilidad y simplicidad, se ha optado por un modelo de scoring de crédito que, aunque menos complejo, ofrece múltiples ventajas. Este modelo permite una interpretación clara de los resultados, lo que facilita la comunicación con los clientes y la toma de decisiones. Además, el estudio de contexto del negocio ha sido fundamental en la toma de esta decisión, asegurando que se alinee con los objetivos y capacidades actuales de la entidad. A pesar de que el modelo de Random Forest presenta métricas superiores en términos de recall, accuracy y ROC, la empresa no está preparada para implementar un modelo de esta naturaleza, ya que requiere una infraestructura técnica más avanzada y una mayor capacitación del personal, lo que podría generar complicaciones en su integración y uso práctico.

9.3 Análisis gobernanza sistemas de información

Actualmente la EFoe no cuenta con un área de analítica, ni de gestión de datos, por lo cual, para identificar las estrategias de gobernabilidad de datos sobre los productos analíticos, se realizó el nivel de madurez enmarcados en los FrameWorks del DAMA, teniendo en cuenta las siguientes áreas de conocimiento:

- Administración: hace referencia a la planificación y gestión de los activos. El objetivo de esta categoría es identificar los responsables de los datos y del manejo de los activos en cada área de negocio, para garantizar las reglas y controles adecuados del manejo de los activos de datos en la EFoe.
- Administración del riesgo: Se entiende como riesgos a la fuga de información, datos errados, violación a la privacidad y seguridad de los activos. En esta categoría se quiere identificar si existe gestión de los riesgos de los activos de datos dentro de cada área de negocio.
- Arquitectura: hace referencia al diseño de los datos estructurados y no estructurados, aplicaciones que habilita la disponibilidad de los datos y su distribución a los usuarios apropiados dentro de cada área de negocio.
- Metadata: Se conoce a la Metadata como una información estructurada sobre datos, que nos los describe sin necesidad de acceder directamente a ellos. En esta Categoría se quiere identificar si existen métodos y las herramientas usadas para crear definiciones semánticas comunes en las áreas de negocio y repositorios de la Metadata.
- Auditoría información de logueo: Se entiende como auditoría de los datos a los procesos para el monitoreo y medición del valor de los datos, mediante el control de los accesos a la

información, generación de reportes, borrado de la información entre otros. En esta categoría se quiere identificar si cada área de negocio tiene metodologías para la auditoría de la información utilizada.

- Calidad: Se refiere a los métodos para medir, mejorar y certificar la calidad e integridad de los datos. En esta categoría se quiere identificar si existen procedimientos para la calidad de los datos en las áreas de negocio.
- Creación de valor: Es el proceso de calificación y cuantificación de los datos para permitir al negocio maximizar el valor creado por los activos. En esta categoría se quiere definir como las áreas del negocio realizan el retorno de la inversión en la recolección, producción y uso de los datos.
- Estructura organizacional: hace referencia a los procesos para el manejo estructurado de los activos, en esta categoría se quiere identificar si cada área de negocio tiene establecida metodologías para el manejo de los datos alineada a las reglas de la organización.
- Política: Es la articulación escrita del comportamiento organizacional deseado. En esta categoría se quiere identificar si existe la creación y formalización de políticas y estándares corporativos alrededor del Gobierno de la Información.

Figura 24 Madurez actual

Nota. Madurez actual de la información en la EFoe. Fuente: elaboración propia

- El nivel de madurez sobre la gestión de los datos en la compañía es un nivel Estable con un valor promedio de 2,6; lo que quiere decir que en este nivel la EFoe ha desarrollado un plan de aseguramiento del valor de los datos y existe compromiso de la alta dirección para orientarse hacia una cultura de la calidad, pero aún no existe una gestión adecuada de la información.
- La Foe aunque no tiene un área de analítica está enfocada a mejorar sus estrategias de datos y tener un nivel de madurez nivel proactivo, por lo que se le recomienda crear el área de analítica de datos, manejo de gestión de información clave, procesos automáticos, herramientas de almacenamiento, arquitectura de datos y gobernanza formal dirigido por la alta dirección.

La EFoe no tiene un área centralizada de analítica, pero tiene definidos procesos de gestión de datos y control de la información en las diferentes áreas. El proyecto empresarial utilizará las siguientes bases de información resumidas en términos de negocio.

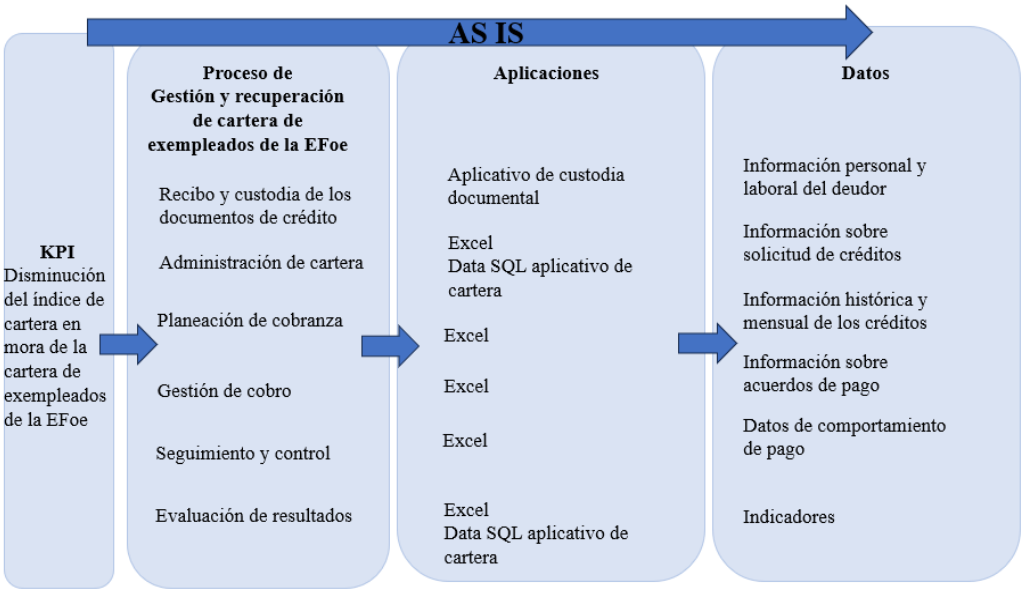
Tabla 4 Descripción de roles involucrados en el proyecto

ID	Rol	Área	Descripción
1	Director de Cartera	Dirección de Cartera	Responsable de la gestión y administración de cartera
2	Profesional de cartera	Dirección de Cartera	Responsable de la gestión y recuperación de la cartera de exempleados
3	Profesional de cartera	Dirección de Cartera	Responsable de la gestión y recuperación de la cartera de empleados
4	Profesional de control de cartera	Dirección de Cartera	Profesional responsable de la planeación y gestión de recuperación de carteras
5	Analista de garantías	Dirección de Cartera	Responsable de la administración y custodia de las garantías de los créditos
6	Secretaría de Cartera	Dirección de Cartera	Responsable de la digitalización de documentos de créditos
7	Jefe de Talento Humano	Jefatura de Talento Humano	Jefe de la jefatura de talento humano
8	Analista de crédito empleados	Jefatura de Talento Humano	Responsable de la recolección y revisión de documentos de crédito
9	Profesional de crédito empleados	Jefatura de Talento Humano	Responsable de la recolección y revisión de documentos de crédito
10	Vicepresidente de operaciones	Vicepresidencia de Operaciones	Dirigir y administrar la vicepresidencia de operaciones
11	Profesional de tecnología	Dirección de Tecnología	Responsable del mantenimiento del aplicativo de cartera
12	Profesional de tecnología	Dirección de Tecnología	Responsable de dar solución a los problemas tecnológicos del área de cartera
13	Líder aplicativo	Proveedor Sonda	Responsable de dar soporte a al aplicativo de cartera

Nota. descripción de los roles involucrados en el proyecto empresarial.
Fuente: elaboración propia.

9.4 Descripción de la situación actual y deseada para gobernanza de datos

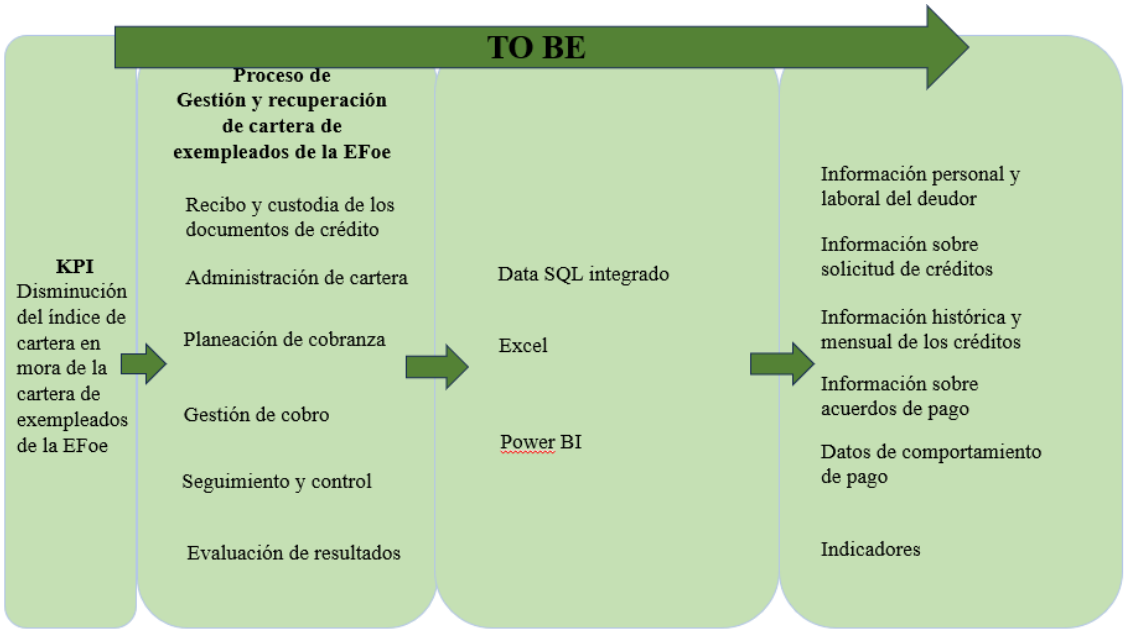
Figura 25 As is



Nota. Descripción actual de la gobernanza de datos en la EFoe.

Fuente: elaboración propia.

Figura 26 To be



Nota. Descripción deseada u objetivo de la gobernanza de datos en la EFoe.

Fuente: elaboración propia.

10 Plan y recomendaciones de implementación y aplicación

Para la correcta implementación del modelo es necesario solicitar al proveedor del aplicativo de cartera en la EFoe la inclusión de un nuevo informe que imprima las siguientes variables:
plazo crédito, plazo residual, endeudamiento, antigüedad cuando se otorgó el crédito, tiempo trabajado, cuentas por cobrar y calificación de riesgo actual, este informe debe actualizarse e imprimirse de manera mensual para con el poder calificar a cada deudor con la tabla de scoring de crédito obtenida.

Una vez realizada la calificación con el modelo de scoring actual se identifica que los deudores que presentan mora están calificados con puntuaciones negativas es por esto que la diferenciación entre deudores buenos y malos es: < a 105 clientes malos > a 105 clientes buenos.

10.1 Próximos Pasos.

Es importante tener en cuenta los próximos pasos para la implementación y ajuste del modelo:

- Estabilidad de variables en producción: Evaluar la estabilidad de las variables una vez que el modelo esté en producción para garantizar que mantengan su relevancia y poder predictivo a lo largo del tiempo.
- Métricas para estimar o calibrar: Establecer métricas claras para identificar cuándo es necesario reestimar o calibrar el modelo. Por ejemplo, si la estabilidad de tres variables clave resulta deficiente en un período de tres meses, se debería planificar una recalibración del modelo.
- Cambios en las variables predictoras: Monitorear si las variables predictoras experimentan cambios significativos en su sentido o proporciones, lo que podría afectar la precisión del modelo, y estar preparado para reestimar el modelo en consecuencia.
- Análisis de sobrevivencia para cobranza preventiva: Realizar un análisis de supervivencia para discriminar en la cobranza preventiva, lo que permitirá desarrollar nuevos modelos de

cobranza que identifique mejor a los clientes en riesgo y optimicen las estrategias de recuperación de deudas.

- Modelo de originación: Desarrollar un modelo de originación para evaluar el riesgo crediticio de nuevos solicitantes, lo que ayudará a tomar decisiones informadas sobre la aprobación o denegación de créditos.
- Alertas y estrategias como proveedor: Implementar alertas y estrategias proactivas para convertirse en un proveedor de soluciones integrales, anticipando las necesidades del cliente y ofreciendo servicios adicionales que agreguen valor a su experiencia.

11 Conclusiones

- El análisis exhaustivo del contexto y los objetivos específicos de la EFoe revela una necesidad imperiosa de optimizar la gestión de recuperación de cartera de exempleados, alineándose con la estrategia general y los indicadores del negocio. La ausencia de una metodología sólida para el estudio de crédito, que permita el análisis de múltiples variables y la determinación de la probabilidad de incumplimiento, ha sido un obstáculo significativo en la gestión eficiente de esta cartera. El indicador de calidad de cartera vencida, que ha alcanzado un preocupante 10.05% en comparación con el 3.91% del promedio de las entidades de crédito del país, es una clara señal de alerta que exige una intervención inmediata. Actualmente, la gestión de recuperación se realiza mediante un plan de cobranza mensual que incluye segmentaciones y utiliza diversos canales de contacto autorizados. Sin embargo, la falta de un modelo de score de incumplimiento limita la capacidad de priorizar a los deudores con mayor riesgo, lo que resulta en un uso ineficiente de recursos. Implementar un modelo predictivo de scoring de recuperación de créditos permitiría focalizar los esfuerzos de recuperación en aquellos deudores con alta probabilidad de incumplimiento, optimizando tiempo y recursos, y mejorando la segmentación.

- La implementación de herramientas de business analytics en la gestión y recuperación de cartera en una entidad financiera se ha demostrado viable y efectiva a través del desarrollo de dos tipos de modelos un modelo de scoring y uno random forest. Estos modelos, diseñados para evaluar y clasificar a los deudores según su probabilidad de incumplimiento, ha permitido identificar con precisión a los deudores buenos y malos. Los resultados

obtenidos evidencian una mejora significativa en la capacidad de la entidad para predecir comportamientos de pago, lo cual es crucial para disminuir el índice de cartera vencida.

- La gobernanza de datos se ha convertido en un elemento esencial para la eficiencia y efectividad en la gestión de cartera dentro de las entidades financieras. En el contexto de la EFoe, la implementación de un nuevo informe que se genere automáticamente desde el aplicativo de gestión de cartera representa un avance significativo hacia una gestión más precisa y proactiva. Este informe incluirá variables críticas como el plazo del crédito, el plazo residual, el nivel de endeudamiento, la antigüedad al momento de otorgar el crédito, el tiempo trabajado, las cuentas por cobrar y la calificación de riesgo. La visualización en un único informe de estas variables permitirá una visión más holística y detallada de cada deudor, facilitando una evaluación más precisa de su situación financiera y su capacidad de pago, la capacidad de generar informes detallados y automatizados desde el aplicativo de gestión de cartera no solo mejora la transparencia y la trazabilidad de los datos, sino que también garantiza que las decisiones se basen en información actualizada y precisa para continuar con la generación del modelo de score.

Referencias bibliográficas

Altamar Perez, N. (16 de agosto de 2023). La cartera vencida en créditos de consumo subió 56,17% y llegó hasta $15,7 billones. *La República*. https://www.larepublica.co/finanzas/la-cartera-vencida-en-creditos-de-consumo-aumento-56-17-y-llego-a-15-7-billones-3681424

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32 https://doi.org/10.1023/A:1010933404324.

Chandra, B., Gupta, M. y Gupta, M.P (2007). *Robust Approach for Estimating Probabilities in Naive-Bayes Classifier*. In International Conference on Pattern Recognition and Machine Intelligence (pp. 11-16).

Chiu, D. (2015). *Machine Learning with R cookbook*. Publishing ltda,

Congreso de la República de Colombia. (10 de julio 2023). Ley 2300: por la cual se establecen medidas que protejan el derecho a la intimidad de los consumidores. *Diario Oficial No. 52452*.

Cramer, J. (2010). *Logit Models from Economics and other Fields.* Cambridge University Press. https://doi.org/10.1017/CBO9780511615412.

Daza Sandoval, L.C. (2015). *Estrategias basadas en el modelo de análisis predictivo árbol de decisión para la mejora del proceso de recaudo de cartera de la línea vehículo particular del banco Davivienda S.A.* [Pontificia Universidad Javeriana].

https://repository.javeriana.edu.co/bitstream/handle/10554/16448/DazaSandovalLauraCarolina2015.pdf?sequence=3

Ghatak, A. (2017). *Machine Learning with R*. Springer Nature.

https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=213990

IBM. (17 de agosto del 2021). *Guía de CRISP-DM de IBM SPSS Modeler*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

IBM. (17 de agosto del 2021). *Máquina de vectores de soporte de Oracle*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=mining-oracle-support-vector-machine-svm

IBM. (22 de noviembre de 2021). SPSS Modeler Subscription. 74 https://www.ibm.com/docs/es/spss-modeler/saas?topic=nodes-clustering-models

Lizares, M. (2017). *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico*. [Universidad Nacional Mayor de San Marcos]. https://unmsm.ent.sirsi.net/client/es_ES/all_libs/search/results?qu=Lizares&te=

Martinez Zapata, D. (2016). *Desarrollo y Validación de Modelo de Scoring de Admisión para Tarjetas de Crédito con metodología de Inferencia de Denegados* [Universidad Carlos III de Madrid]. https://documentacion.fundacionmapfre.org/documentacion/publico/pt/bib/158007.do

Mendez Anaya, W.A & Galvis Jurado, M.A. (2019). *Modelo de cálculo de la Probabilidad de Recuperación de la Cartera Castigada en la Agencia de Financiera Comultrasan en San Gil* [Universidad Autónoma de Bucaramanga]. https://repository.unab.edu.co/bitstream/handle/20.500.12749/14690/2019_Tesis_Galvis_Jurado_Mario_Alexander.pdf?sequence=1

Nieto, S. (2010). *Crédito al consume: La estadística aplicada a un problema de riesgo crediticio.* [Universidad Autónoma Metropolitana]. http://www.academia.edu/8454174/Proyecto_de_Tesis_Cr%C3%A9dito_al_Consumo_La_Estad%C3%ADstica_aplicada_a_un_problema_de_Riesgo_Crediticio

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Superintendencia Financiera de Colombia. (2021). *Circular Externa 018 de 2021, Capítulo XXXI Sistema Integral de Administración de Riesgos SIAR. Parte I Generalidades*. https://www.superfinanciera.gov.co/loader.php?lServicio=Tools2&lTipo=descargas&lFuncion=descargar&idFile=1055798

Superintendencia Financiera de Colombia. (2024). *Sobre la protección de datos personales*. https://www.sic.gov.co/content/sobre-la-protecci%C3%B3n-de-datos-personales#:~:text=Dato%20Privado%3A,indebido%20puede%20generar%20su%20discriminaci%C3%B3n.

Superintendencia Financiera de Colombia. (2023). *Buscador de terminos*. https://www.superfinanciera.gov.co/glosario/buscar/www.superfinanciera.gov.co?q=cartera

Tabladillo, M. (27 de febrero de 2024). The team data science process lifecycle. *Learn Microsoft.* https://learn.microsoft.com/en-us/azure/architecture/data-science-process/lifecycle

The Circus (13 de marzo de 2023). Riesgo de crédito: qué son y cuales hay. *The Circus*. https://www.santanderconsumer.es/blog/post/riesgo-de-credito-o-crediticio-que-son-y-cuales-hay#:~:text=El%20riesgo%20de%20cr%C3%A9dito%20o,no%20recupere%20el%20dinero%20prestado.

Villamil Bahamon, R. (2013). *Modelo predictivo neuronal para la evaluación del riesgo de crédito*. [Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/bitstream/handle/unal/52246/08901050.2013.pdf?sequence=1&isAllowed=y