Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Algoritmos de Machine Learning para la priorización de clientes en campañas comerciales en un call center de ventas, MF Asesoría y Consultoría 2024

TRABAJO DE SUFICIENCIA PROFESIONAL
Para optar el Título Profesional de Licenciado en Estadística

AUTOR

Luis Daniel VILLACORTA TITO

ASESOR

Roberto Carlos FIESTAS FLORES

Lima, Perú

2025



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

https://creativecommons.org/licenses/by-nc-sa/4.0/

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Villacorta, L. (2025). *Algoritmos de Machine Learning para la priorización de clientes en campañas comerciales en un call center de ventas, MF Asesoría y Consultoría 2024*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Luis Daniel Villacorta Tito
Tipo de documento de identidad	DNI
Número de documento de identidad	75877654
URL de ORCID	https://orcid.org/0009-0008-4641-3050
Datos de asesor	
Nombres y apellidos	Roberto Carlos Fiestas Flores
Tipo de documento de identidad	DNI
Número de documento de identidad	16744141
URL de ORCID	https://orcid.org/0000-0002-5582-0124
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Oscar Antonio Robles Villanueva
Tipo de documento de identidad	DNI
Número de documento de identidad	32762171
Miembro del jurado 1	
Nombres y apellidos	Bernardo Céspedes Panduro
Tipo de documento de identidad	DNI
Número de documento de identidad	17615559
Datos de investigación	
Línea de investigación	A.3.2.6 Análisis de Datos y Modelamiento de Problemas de la Sociedad

Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	Calle Fermin Tanguis 151, La Victoria, Lima 15034 País: Perú Departamento: Lima Provincia: Lima Ciudad: Lima Distrito: La Victoria Latitud: -12.0809258 Longitud: -77.0118495
Año o rango de años en que se realizó la investigación	2025
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

ACTA DE SUSTENTACIÓN DE TRABAJO DE SUFICIENCIA PROFESIONAL PARA LA OBTENCIÓN DEL TÍTULO PROFESIONAL DE LICENCIADO EN ESTADÍSTICA (PROGRAMA DE TITULACIÓN PROFESIONAL 2025)

En la UNMSM – Ciudad Universitaria – Facultad de Ciencias Matemáticas, siendo las 18:40 horas del martes 3 de junio del 2025, se reunieron los docentes designados como Miembros del Jurado Evaluador (PROGRAMA DE TITULACIÓN PROFESIONAL 2025): Dr. Oscar Antonio Robles Villanueva (PRESIDENTE), Dr. Bernardo Céspedes Panduro (MIEMBRO) y el Mg. Roberto Carlos Fiestas Flores, (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: **"ALGORITMOS DE MACHINE LEARNING PARA LA PRIORIZACIÓN DE CLIENTES EN CAMPAÑAS COMERCIALES EN UN CALL CENTER DE VENTAS, MF ASESORÍA Y CONSULTORÍA 2024"**, presentado por el señor **Bachiller LUIS DANIEL VILLACORTA TITO**, para optar el Título Profesional de Licenciado en Estadística.

Luego de la exposición del Trabajo de Suficiencia Profesional, el Presidente invitó al expositor a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los Miembros del Jurado Evaluador, el expositor mereció la aprobación aprobado con mención honrosa con un calificativo promedio de 16 (dieciseis)

A continuación, los Miembros del Jurado Evaluador dan manifiesto que el participante **Bachiller LUIS DANIEL VILLACORTA TITO**, en vista de haber aprobado la sustentación de su Trabajo de Suficiencia Profesional, será propuesto para que se le otorgue el Título Profesional de Licenciado en Estadística.

Siendo las 19:20 horas se levantó la sesión firmando para constancia la presente Acta.

Dr. Oscar Antonio Robles Villanueva
PRESIDENTE

Dr. Bernardo Céspedes Panduro
MIEMBRO

Mg. Roberto Carlos Fiestas Flores
MIEMBRO ASESOR

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú, Decana de América

CERTIFICADO DE SIMILITUD

Yo, ROBERTO CARLOS FIESTAS FLORES en mi condición de asesor acreditado con Dictamen N° 000010-2025-EPEST-FCM/UNMSM del trabajo de suficiencia profesional, cuyo título es: **Algoritmos de Machine Learning para la Priorización de Clientes en Campañas Comerciales en un Call Center de Ventas, MF Asesoría y Consultoría 2024**, presentado por el bachiller Luis Daniel VILLACORTA TITO para optar el título de Licenciado en Estadística, CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el 2% de similitud, nivel PERMITIDO para continuar con los trámites correspondientes y para su publicación en el repositorio institucional.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título de Licenciado en estadística.

Lima, 16 de junio de 2025.



Firma del Asesor _____

Nombres y Apellidos del asesor: Fiestas Flores, Roberto Carlos

DNI 16744141

Resumen

El presente trabajo de suficiencia profesional tuvo como objetivo desarrollar e implementar un modelo de machine learning que permita priorizar clientes en una campaña comercial de ventas telefónicas de productos financieros. La actividad se realizó en MF Asesoría y Consultoría S.A.C., empresa especializada en la gestión de fuerzas de venta para entidades del sector financiero. El problema identificado fue la asignación subóptima del esfuerzo comercial debido al uso exclusivo de juicio experto. Se aplicó una metodología basada en CRISP-DM, incluyendo etapas de exploración, limpieza y transformación de datos, así como la selección y evaluación de distintos algoritmos de clasificación. Tras comparar más de 10 modelos y ajustar hiperparámetros mediante RandomGridSearch, se seleccionó Linear Discriminant Analysis (LDA) por su rendimiento y capacidad de generar probabilidades. El modelo fue integrado en un pipeline y desplegado en producción. Los resultados evidencian un incremento en la productividad y efectividad comercial, con un AUC de 0.713 y un Gini de 0.426. Se concluye que el enfoque basado en datos permite una asignación más eficiente de recursos y tiene potencial para ser replicado en otras campañas.

Palabras clave: Machine learning, priorización de clientes, clasificación, call center, CRISP-DM.

Abstract

This professional sufficiency project aimed to develop and implement a machine learning model to prioritize clients in a telemarketing campaign offering financial products. The activity was carried out at MF Asesoría y Consultoría S.A.C., a Peruvian company specialized in sales force management. The main issue was the suboptimal allocation of commercial effort due to the exclusive use of expert judgment. A CRISP-DM-based methodology was applied, including data exploration, cleaning, transformation, feature selection, and model evaluation. After testing over 10 algorithms and fine-tuning via RandomGridSearch, Linear Discriminant Analysis (LDA) was selected for its predictive performance and ability to return probabilities. The model was integrated into a modular pipeline and deployed in production. Results show a significant improvement in commercial effectiveness and productivity, with an AUC of 0.713 and a Gini coefficient of 0.426. The study concludes that data-driven approaches allow for a more efficient allocation of resources and have high potential for replication in future campaigns.

Keywords: Machine learning, client prioritization, classification, call center, CRISP-DM.

Índice

I. Introducción 5

II. Información del lugar donde se desarrolló la actividad 6

2.1. Institución donde se desarrolló la actividad 6

2.2. Periodo de duración de la actividad 6

2.3. Finalidad y objetivo de la entidad 7

2.4. Razón social 7

2.5. Dirección postal 7

2.6. Correo electrónico del profesional a cargo 7

III. Descripción de la actividad 8

3.1. Organigrama de la entidad 8

3.2. Finalidad y objetivos de la actividad 9

3.3. Problemática 9

3.4. Metodología, Procedimientos 10

3.5. Resultados de la actividad 19

IV. Conclusiones 28

4.1. Justificación 29

4.2. Metodología aplicada 29

4.3. Descripción de la implementación 30

V. Recomendaciones 32

VI. Bibliografía 34

VII. ANEXOS e Ilustraciones 36

I. Introducción

En la actualidad, las empresas que operan mediante call centers enfrentan el reto de maximizar su productividad comercial en un entorno altamente competitivo. Este desafío se intensifica cuando los productos ofrecidos requieren de una presentación activa y efectiva, como es el caso de los préstamos de libre disposición. Tradicionalmente, la asignación de esfuerzos se ha guiado por la experiencia acumulada de supervisores y ejecutivos, lo cual introduce un grado importante de subjetividad y limita la capacidad para identificar patrones complejos en el comportamiento del cliente.

Frente a esta problemática, el presente informe documenta el desarrollo de un sistema de priorización de clientes basado en algoritmos de clasificación, aplicado en el contexto de una campaña comercial real. La actividad se realizó en MF Asesoría y Consultoría S.A.C. y tuvo como objetivo principal construir una herramienta que permita predecir la probabilidad de aceptación de cada cliente, con el fin de ordenar la gestión comercial de forma más eficiente.

El informe está estructurado en siete secciones. Primero, se presenta el contexto de la actividad y la problemática identificada. Luego, se describe la metodología aplicada, fundamentada en CRISP-DM. A continuación, se detallan los procedimientos de selección de variables, modelado, validación y generación de scores. Posteriormente, se exponen los resultados obtenidos en términos operativos y técnicos. Finalmente, se presentan las conclusiones y recomendaciones para futuras implementaciones.

II. Información del lugar donde se desarrolló la actividad

2.1. Institución donde se desarrolló la actividad

La actividad profesional se llevó a cabo en **MF Asesoría y Consultoría**, una empresa peruana fundada en 2003 que ofrece soluciones comerciales a diversas organizaciones mediante la formación de equipos de ventas exclusivos, especializados y capacitados para cada cliente.

Inicialmente enfocada en el sector financiero, MF Asesoría y Consultoría ha ampliado su cartera de clientes para incluir organizaciones no gubernamentales como UNICEF y empresas del sector salud como Auna Salud. Esta diversificación refleja su capacidad para adaptarse a diferentes mercados y ofrecer servicios personalizados que satisfacen las necesidades específicas de cada cliente.

Entre los servicios que brinda la empresa se encuentran la gestión de fuerzas de ventas a través de call centers, televentas, ventas receptivas y ventas de campo; la evaluación, reclutamiento y selección de personal comercial; el servicio de vendedores; y el planeamiento comercial para dimensionar las fuerzas de ventas según el mercado.

2.2. Periodo de duración de la actividad

La actividad profesional se desarrolló en el periodo comprendido entre los meses de enero de 2024 y julio de 2024. Durante este tiempo, se realizaron diversas tareas vinculadas al análisis de datos, desarrollo de modelos predictivos y mejora de procesos dentro del entorno operativo de la empresa, permitiendo aplicar conocimientos técnicos en un contexto real y alineado a los objetivos institucionales.

2.3. Finalidad y objetivo de la entidad

La finalidad de MF Asesoría y Consultoría es brindar soluciones comerciales integrales a empresas de diversos sectores mediante la conformación de equipos de ventas especializados y la implementación de estrategias comerciales efectivas. Su objetivo principal es ofrecer servicios como la gestión de fuerza de ventas, el reclutamiento y selección de personal comercial, la administración de planillas y el planeamiento estratégico, con el propósito de optimizar los procesos comerciales de sus clientes, incrementar su eficiencia operativa y potenciar sus resultados en el mercado.

2.4. Razón social

La razón social de la entidad donde se desarrolló la actividad es MF Asesoría y Consultoría S.A.C., registrada legalmente como una sociedad anónima cerrada en el Perú. Esta denominación representa su identidad jurídica ante las autoridades y entidades con las que mantiene relaciones comerciales y contractuales.

2.5. Dirección postal

La dirección legal de MF Asesoría y Consultoría S.A.C. es Calle Fermín Tangüis 145, Oficina 903, distrito de La Victoria, Lima 15034, Perú. Desde esta sede, la empresa coordina sus operaciones administrativas y comerciales a nivel nacional.

2.6. Correo electrónico del profesional a cargo

El profesional responsable del desarrollo de la actividad puede ser contactado a través del siguiente correo electrónico personal: luis.villacorta.tito@gmail.com.

III. Descripción de la actividad

3.1. Organigrama de la entidad

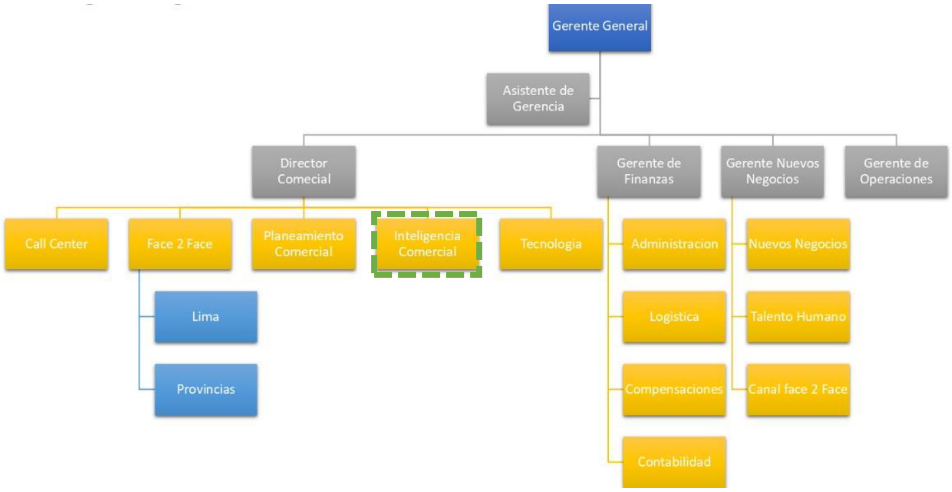
La actividad fue desarrollada desde el área de Inteligencia Comercial, parte de la estructura de la Dirección Comercial, ver Figura 1. Esta unidad se encarga del análisis y aprovechamiento de datos para apoyar decisiones estratégicas y optimizar campañas comerciales, con énfasis en la priorización de recursos.

Según el organigrama comercial vigente en 2024, la Dirección Comercial también comprende áreas como Call Center, Planeamiento Comercial, Tecnología y Face 2 Face, abarcando funciones clave de contacto con clientes y planificación operativa. La Gerencia General, apoyada por la Asistencia de Gerencia, supervisa transversalmente a todas las gerencias, asegurando la alineación estratégica.

En este contexto, el desarrollo del sistema de priorización de clientes mediante técnicas de machine learning fue una iniciativa de Inteligencia Comercial, con impacto directo en la eficiencia del Call Center y potencial de extensión hacia otros canales como Face 2 Face.

Figura 1

Organigrama comercial de la empresa en 2024



Nota. Elaboración propia a partir de la estructura organizacional vigente.

3.2. Finalidad y objetivos de la actividad

La actividad se desarrolló en respuesta a una necesidad planteada por la Gerencia Comercial y la Jefatura de Unidad de la Campaña, con el propósito de mejorar la eficiencia en la gestión de clientes durante las campañas comerciales. La finalidad principal fue implementar un sistema de priorización que permitiera enfocar los esfuerzos del equipo de ventas en los clientes con mayor probabilidad de conversión, optimizando así los recursos disponibles y elevando los indicadores de efectividad comercial.

Los objetivos específicos fueron:

- Explorar y analizar las bases de datos históricas de campañas comerciales para identificar variables relevantes asociadas al cierre de ventas.
- Preparar y procesar la información para construir un conjunto de datos adecuado para tareas de clasificación binaria.
- Evaluar el desempeño de diferentes algoritmos de machine learning en la predicción de la probabilidad de conversión.
- Seleccionar el modelo con mejor capacidad de clasificación, según métricas como el AUC, F1-score y precisión.
- Implementar una propuesta de priorización de clientes basada en las predicciones del modelo seleccionado.

3.3. Problemática

En la campaña comercial gestionada para una entidad financiera que ofrece productos de préstamo de libre disposición (PLD), se trabaja mensualmente con bases de datos amplias de clientes potenciales. La priorización de estos clientes se realiza actualmente en base al juicio experto de los supervisores y al conocimiento acumulado de campañas anteriores. Si

bien esta metodología ha permitido cierto nivel de eficiencia, presenta limitaciones importantes.

El principal problema radica en que los criterios utilizados no siempre logran capturar la complejidad de los patrones que influyen en la aceptación del producto. Existen combinaciones de variables y comportamientos que pueden pasar desapercibidos, lo que dificulta identificar con precisión a los clientes con mayor probabilidad de conversión.

Esto genera una asignación subóptima del esfuerzo comercial, donde clientes con alta disposición a aceptar el producto pueden no ser gestionados a tiempo o recibir menos intentos, mientras que se dedican recursos a perfiles de menor potencial. Esta situación impacta negativamente en los indicadores de eficiencia, tasa de conversión y rentabilidad de la campaña.

Como señalan Davenport y Harris (2007), en entornos operativos complejos, las organizaciones que basan sus decisiones únicamente en juicio experto tienden a obtener resultados subóptimos frente a aquellas que incorporan modelos analíticos para optimizar sus procesos.

Ante esta situación, se plantea la necesidad de complementar la experiencia del equipo con una herramienta basada en modelos de machine learning que permita detectar patrones complejos y establecer una priorización objetiva y efectiva de la base de clientes a gestionar.

3.4. Metodología, Procedimientos

Este apartado describe el enfoque metodológico y las etapas ejecutadas para el desarrollo del sistema de clasificación y segmentación de clientes en el contexto de la campaña PLD de una entidad financiera.

Comprensión del negocio y levantamiento de requerimientos:

Se realizaron reuniones con la Gerencia Comercial y la Jefatura de Unidad de la Campaña PLD con el objetivo de comprender la lógica operativa del proceso, los criterios actuales de priorización de clientes y los principales indicadores de conversión utilizados por el área comercial. Esta interacción permitió identificar los objetivos del negocio y traducirlos en objetivos analíticos concretos para el desarrollo del sistema de clasificación y segmentación de clientes.

Desde el enfoque metodológico, esta etapa se alinea con la fase de "Comprensión del negocio" de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), considerada un estándar en proyectos de ciencia de datos. Según Chapman et al. (2000), esta fase inicial es crucial para asegurar que los desarrollos analíticos estén alineados con las necesidades reales de la organización, permitiendo definir correctamente las metas del análisis, los criterios de éxito y las restricciones operativas.

Recolección y exploración de datos:

Se consolidaron las bases de datos históricas provenientes de la gestión de llamadas, la tipificación de contactos y los registros de desembolsos efectivos, a fin de contar con una visión completa del comportamiento de los clientes durante la campaña. Posteriormente, se realizó un análisis exploratorio de datos que permitió evaluar la calidad de la información disponible, identificar valores faltantes o inconsistencias, y detectar patrones iniciales útiles para el proceso de modelado.

Esta fase corresponde al enfoque del Análisis Exploratorio de Datos (EDA) propuesto por Tukey (1977), el cual promueve la exploración visual y estadística de los datos antes de aplicar técnicas formales de modelado. El EDA permite al analista descubrir relaciones no

evidentes, guiar la generación de hipótesis y evitar errores derivados de una comprensión limitada de los datos. Asimismo, se consideran aportes metodológicos de Han et al. (2011) respecto a la preparación y entendimiento del dataset como paso previo fundamental dentro de los procesos de minería de datos.

Preparación y limpieza de datos:

Con el objetivo de estructurar un conjunto de datos apto para el entrenamiento de modelos predictivos, se ejecutaron diversas tareas de transformación y depuración. Estas incluyeron la imputación de valores faltantes, la codificación de variables categóricas, la detección y tratamiento de valores atípicos, así como la creación de nuevas variables derivadas que pudieran capturar relaciones no explícitas entre las variables originales.

Desde el enfoque estadístico, la imputación se justifica en la necesidad de evitar sesgos derivados del manejo inapropiado de datos ausentes. Little y Rubin (2002) establecen tres mecanismos fundamentales que explican el patrón de ausencia:

- **Missing Completely at Random (MCAR)**: cuando la ausencia es completamente independiente de cualquier variable observada o no observada.
- **Missing at Random (MAR)**: cuando la probabilidad de que falte un dato depende solo de variables observadas.
- **Missing Not at Random (MNAR)**: cuando la ausencia está relacionada directamente con el valor faltante en sí.

Estas categorías son claves para seleccionar el método de imputación apropiado, ya que los métodos clásicos, como la imputación por media o regresión, solo ofrecen estimaciones no sesgadas si se cumple al menos el supuesto MAR. En este análisis, se asumió

dicho supuesto como punto de partida, lo cual permitió el uso de técnicas condicionales de imputación sin comprometer la validez estadística del proceso.

En cuanto a la detección de valores atípicos, se aplicaron procedimientos basados en reglas estadísticas como el rango intercuartílico (IQR) y el método de puntuaciones z. El IQR evalúa la dispersión entre el primer y el tercer cuartil para establecer un rango aceptable. Una observación se considera atípica si está fuera del intervalo:

$$Límite\ inferior = Q1 - 1.5 \times IQR$$

$$Límite\ superior = Q3 + 1.5 \times IQR$$

Por su parte, el z-score mide la distancia de una observación respecto a la media en términos de desviaciones estándar. Una observación se considera atípica si su valor absoluto es mayor a 3:

$$z_i = \frac{x_i - \bar{x}}{S}$$

Ambos métodos son recomendados por Iglewicz y Hoaglin (1993) y se utilizaron según la distribución y escala de las variables analizadas, permitiendo preservar la calidad del análisis sin eliminar observaciones valiosas de forma injustificada.

Las variables categóricas fueron transformadas mediante técnicas como label encoding y one-hot encoding (ver Tabla 1), con el fin de representar información cualitativa en una estructura numérica adecuada para su uso en algoritmos de aprendizaje automático. El label encoding asigna un número entero a cada categoría, mientras que el one-hot encoding crea una columna binaria por cada categoría, codificando la presencia o ausencia.

Tabla 1

Comparación de técnicas de codificación categórica y sus supuestos estadísticos

Técnica	Descripción	Supuesto / Implicancia estadística	Casos recomendados
Label Encoding	Asigna un número entero a cada categoría (ej. A=0, B=1, C=2)	Supone orden implícito; puede inducir relaciones falsas	Modelos basados en árboles (no sensibles al orden)
One-Hot Encoding	Crea una columna binaria por cada categoría	Evita supuestos de orden; puede aumentar dimensionalidad	Modelos lineales, distancia o redes neuronales

Nota. Elaboración propia.

En este trabajo se optó por one-hot encoding para modelos sensibles a la escala (como la regresión logística), y label encoding para algoritmos basados en árboles (como Random Forest o XGBoost), dado que estos no interpretan el orden numérico como una relación ordinal.

Finalmente, se realizó ingeniería de variables, basada en la creación de nuevas características derivadas de combinaciones, transformaciones o segmentaciones de las variables originales. Esta etapa permitió representar relaciones no lineales o patrones específicos que enriquecieron el poder predictivo del conjunto de datos. Este proceso es ampliamente recomendado en proyectos de minería de datos por autores como Han, Kamber y Pei (2011), quienes destacan su rol clave en el desempeño de los modelos.

Selección de variables:

Con el objetivo de mejorar la capacidad predictiva de los modelos y reducir la complejidad computacional, se implementó un proceso de selección de variables que permitió identificar los atributos con mayor relevancia explicativa. Esta etapa resulta fundamental en la construcción de modelos robustos, ya que reduce el riesgo de sobreajuste, mejora la interpretabilidad y disminuye los tiempos de entrenamiento, favoreciendo la eficiencia del proceso de modelado.

Desde el enfoque estadístico, se evaluó la presencia de multicolinealidad mediante el cálculo del coeficiente de correlación de Pearson, el cual mide la relación lineal entre dos variables continuas. Este coeficiente se define como:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Valores absolutos altos de $\boldsymbol{r}$, cercanos a 1, indican redundancia entre variables, lo cual afecta directamente a modelos como la regresión lineal y logística, generando varianzas infladas en los estimadores de los coeficientes. Para mitigar estos efectos, se revisó la matriz de correlaciones entre los predictores, eliminando aquellos con correlaciones superiores a un umbral predeterminado ($\boldsymbol{r} > 0.8$)

Desde el enfoque de aprendizaje automático, se utilizaron técnicas embebidas y wrapper que permiten evaluar la importancia de las variables en función de su impacto directo sobre el desempeño del modelo. En el caso de modelos como árboles de decisión y Random Forest, se utilizó la importancia de variables basada en la reducción de impureza. Esta impureza puede calcularse mediante métricas como el índice de Gini:

$$Gini(t) = 1 - \sum_{i=1}^{C} {p_i}^2$$

o la entropía de Shannon:

$$H(t) = -\sum_{i=1}^{C} p_i log_2 p_i$$

Donde p_i representa la proporción de observaciones pertenecientes a la clase i en el nodo t. Ambas métricas buscan cuantificar cuán homogéneo es un nodo respecto a la variable objetivo. Las variables que generan mayores reducciones de impureza a lo largo del árbol

reciben mayor puntuación de importancia. En Random Forest, esta importancia se promedia entre todos los árboles del conjunto, entregando una medida robusta de la relevancia de cada predictor.

También se aplicó la técnica Recursive Feature Elimination (RFE), un método wrapper que selecciona variables de manera iterativa. Se parte de un conjunto completo de predictores, se entrena un modelo base (por ejemplo, regresión logística), se evalúa la importancia de cada variable y se elimina aquella con menor peso relativo. El proceso se repite hasta alcanzar un número óptimo de variables, validado con métricas como la exactitud, el AUC o el F1-score.

Adicionalmente, se utilizó el modelo penalizado Lasso (Least Absolute Shrinkage and Selection Operator), que incorpora un término de regularización L_1 en la función de pérdida:

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Esta penalización promueve soluciones esparsas, reduciendo a cero los coeficientes de las variables menos relevantes, lo cual constituye una forma eficiente de selección automática de variables. Este enfoque es especialmente útil cuando se trabaja con conjuntos de datos de alta dimensionalidad o con variables correlacionadas, ya que mejora la estabilidad y generalización del modelo final.

Tal como argumentan Guyon y Elisseeff (2003), una adecuada selección de características incrementa la interpretabilidad del modelo, mejora su capacidad predictiva y reduce su complejidad sin comprometer su desempeño. Por tanto, esta etapa representa un componente esencial dentro del flujo metodológico del presente análisis.

Construcción y evaluación de modelos:

Con el objetivo de identificar el modelo con mayor capacidad predictiva, se entrenaron distintos algoritmos de clasificación supervisada. Entre ellos destacan los árboles de decisión (Breiman, 1984), Random Forest (Breiman, 2001) y XGBoost (Chen & Guestrin, 2016), los cuales han demostrado un desempeño competitivo en problemas de clasificación binaria con datos estructurados.

El proceso de entrenamiento se realizó bajo un esquema de validación cruzada estratificada, que permitió estimar el desempeño de los modelos de forma robusta y evitar el sobreajuste. Esta técnica consiste en dividir el conjunto de datos en k subconjuntos (folds), entrenando el modelo en k-1 de ellos y evaluando en el restante, repitiendo el proceso k veces.

Para la evaluación, se utilizaron métricas comúnmente empleadas en contextos con clases desbalanceadas:

- Precisión (Accuracy):

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

- F1-score:

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

- Área bajo la curva ROC (AUC):

  Esta métrica evalúa la capacidad del modelo para distinguir entre clases, siendo especialmente útil cuando el conjunto está desbalanceado. Se interpreta como la probabilidad de que el modelo asigne una mayor puntuación a una observación positiva que a una negativa.

Además, se monitoreó el comportamiento de los modelos sobre los conjuntos de entrenamiento y prueba para detectar posibles problemas de overfitting o underfitting, analizando el trade-off entre sesgo y varianza en cada algoritmo.

Selección del modelo óptimo:

Tras evaluar el desempeño de diversos algoritmos de clasificación, se procedió a seleccionar el modelo con mayor capacidad predictiva, considerando tanto su rendimiento promedio en métricas clave (AUC, F1-score, precisión), como su estabilidad a través de los diferentes folds de validación cruzada.

Esta etapa se enmarca dentro del proceso de optimización de modelos propuesto por Kuhn y Johnson (2013), quienes destacan la importancia de balancear el rendimiento predictivo con la consistencia del modelo frente a nuevas muestras de datos. Un modelo con alta varianza entre conjuntos de validación puede evidenciar un sobreajuste, mientras que uno estable y preciso representa una solución más confiable para ser implementada operativamente.

El modelo seleccionado fue utilizado como base para calcular las probabilidades individuales por cliente y construir una estrategia de priorización basada en esos scores.

Generación de scores de probabilidad:

Una vez identificado el modelo con mejor rendimiento, se procedió a su aplicación sobre una nueva base de datos correspondiente a clientes por gestionar. El modelo devolvió una probabilidad individual por cliente, indicando su propensión a aceptar el producto ofertado. Estas probabilidades, también conocidas como scores de propensión, constituyen una herramienta esencial para orientar la estrategia de contacto en campañas comerciales.

Desde el enfoque estadístico y de ciencia de datos, los algoritmos de clasificación permiten obtener no solo una clase predicha, sino también la probabilidad asociada a cada clase. Esta salida probabilística se interpreta como la confianza del modelo en la predicción, y resulta particularmente útil para ordenar y segmentar clientes en función de su potencial (Provost & Fawcett, 2013).

Elaboración del ranking de priorización:

Con base en los scores de probabilidad obtenidos, se construyó un ranking de clientes, ordenándolos de mayor a menor según su probabilidad estimada de aceptación. Este ranking fue entregado al equipo de gestión para priorizar sus esfuerzos de contacto, optimizando así los recursos disponibles y aumentando la eficiencia de la campaña.

Esta estrategia está respaldada por prácticas de marketing dirigido y scoring crediticio, donde los clientes son priorizados según su propensión a responder positivamente ante una acción comercial (Hand & Henley, 1997). La priorización basada en scores permite focalizar el trabajo comercial en los segmentos más rentables, maximizando el retorno de la campaña sin incrementar el esfuerzo operativo.

3.5. Resultados de la actividad

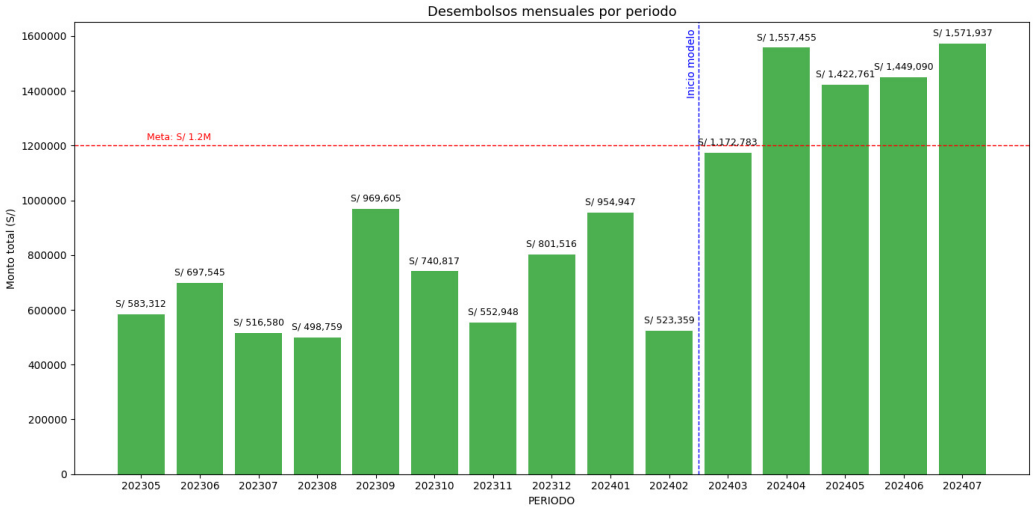
Como parte de la presente actividad, se analizó la evolución de los resultados comerciales desde mayo de 2023 (202305) hasta julio de 2024 (202407), abarcando tanto el periodo previo como posterior a la implementación del modelo de machine learning. El modelo fue puesto en producción en marzo de 2024 (202403), con el objetivo de priorizar de manera más eficiente a los clientes con alta probabilidad de aceptación del producto ofrecido. La meta de desembolso establecida por la campaña fue de S/ 1.2 millones mensuales.

Antes de la implementación del modelo, los niveles de producción mensual presentaban alta variabilidad y, en general, no alcanzaban la meta establecida. No obstante, a

partir del despliegue del modelo, se observa un cambio significativo en la tendencia. Desde el periodo 202403 en adelante, todos los meses superaron la meta de S/ 1.2 millones, alcanzando máximos de hasta S/ 1.57 millones (ver Figura 2). Este comportamiento sugiere una mejora en la efectividad de la asignación comercial, facilitada por el uso del modelo predictivo.

Figura 2

Evolución mensual del monto desembolsado

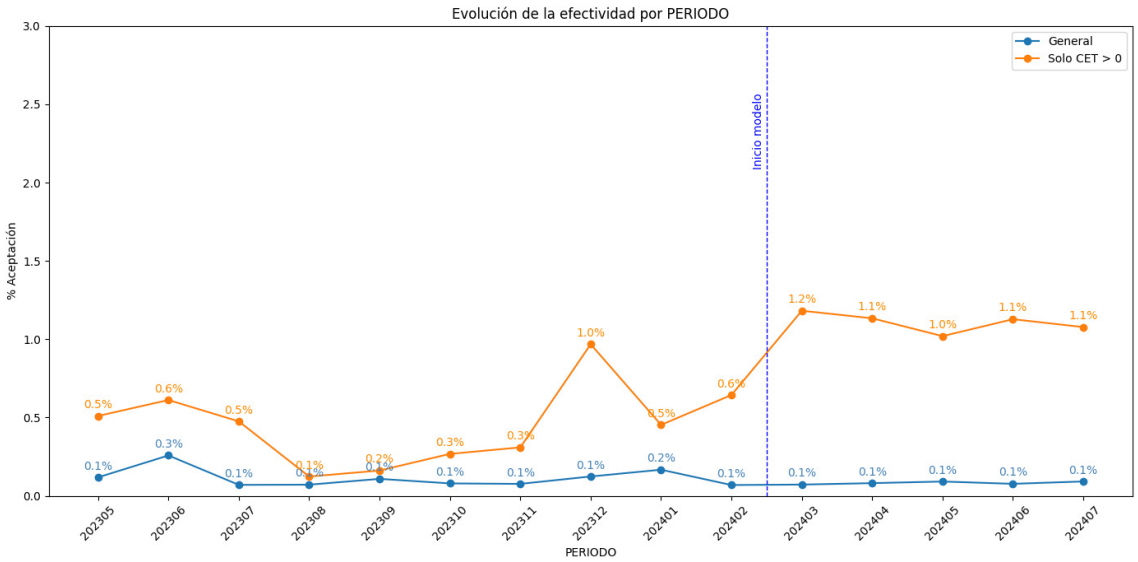


Nota. Elaboración propia basado en datos internos.

Una métrica clave para evaluar el rendimiento comercial es la efectividad, entendida como el porcentaje de clientes que aceptaron la oferta frente al total de registros gestionados. En ese sentido, se diferenciaron dos grupos: el universo total de clientes y aquellos con interacción efectiva (cet > 0), es decir, quienes realmente recibieron la oferta comercial. Esta última categoría resulta más representativa, ya que considera únicamente a los contactos en los que el ejecutivo logró presentar la propuesta.

Figura 3

Efectividad mensual comparada: general vs. clientes contactados



Nota. Elaboración propia basado en datos internos.

Tal como se muestra en la Figura 3, la efectividad general siempre se mantuvo por debajo del 0.2%, debido a la alta proporción de clientes no contactados. En contraste, la efectividad entre los clientes con contacto (cet > 0) muestra un crecimiento marcado a partir del periodo 202403, momento en el que el modelo fue implementado. Desde ese punto, la tasa de aceptación supera sostenidamente el 1%, duplicando o incluso triplicando los niveles previos, lo cual refleja una mayor precisión en la asignación del esfuerzo comercial hacia perfiles con alta probabilidad de conversión.

Este resultado confirma que el modelo no solo incrementó la producción total, sino que también mejoró la **calidad del contacto**, priorizando correctamente a los clientes con mayor propensión a aceptar la oferta financiera.

El modelo fue evaluado utilizando métricas estándar de clasificación en los conjuntos de entrenamiento y prueba. Tal como se muestra en la Tabla 2, el valor del AUC alcanzó 0.736 en entrenamiento y 0.713 en prueba, lo cual refleja una buena capacidad del modelo para discriminar entre clientes propensos y no propensos a aceptar la oferta.

Tabla 2

Métricas globales del modelo de clasificación en los conjuntos de entrenamiento y prueba

Métrica	Train	Test
Accuracy	0.723	0.722
Precision	0.009	0.008
Recall	0.750	0.705
F1-score	0.017	0.016
AUC	0.736	0.713
Gini	0.472	0.426

Nota. Elaboración propia.

La métrica de recall fue relativamente alta (0.750 en entrenamiento y 0.705 en prueba), lo que indica que el modelo logra identificar una proporción significativa de los casos positivos reales. Esta característica es especialmente relevante en campañas comerciales, donde omitir a un cliente con alta probabilidad de conversión representa una pérdida directa de oportunidad.

Por otro lado, la precisión y el F1-score presentan valores bajos, lo cual es esperable en contextos con fuerte desbalance de clases, como en este caso. No obstante, desde un enfoque operativo, esta baja precisión no representa una desventaja crítica. Esto se debe a que la estrategia comercial del cliente (el banco) implica gestionar de todas formas la base total asignada. Por tanto, el objetivo del modelo no es excluir casos, sino ordenarlos por prioridad, de modo que los ejecutivos comiencen gestionando a los clientes con mayor probabilidad de aceptación.

En ese sentido, el modelo cumple satisfactoriamente su propósito: facilitar una priorización eficiente que maximice la conversión en los primeros intentos, optimizando el uso del tiempo y los recursos humanos de la campaña.

Durante el desarrollo del modelo se evaluaron inicialmente 19 variables predictoras, todas relacionadas al perfil financiero y comercial del cliente. Con el objetivo de seleccionar las variables más relevantes y reducir la complejidad del modelo, se aplicaron múltiples métodos de selección de variables, entre ellos:

- Importancia por RandomForest, XGBoost, y LightGBM
- Métodos wrapper como RFE (Recursive Feature Elimination)
- Métodos penalizados como Lasso
- Selección basada en SelectKBest, Boruta
- Importancia estructural vía SelectFromModel

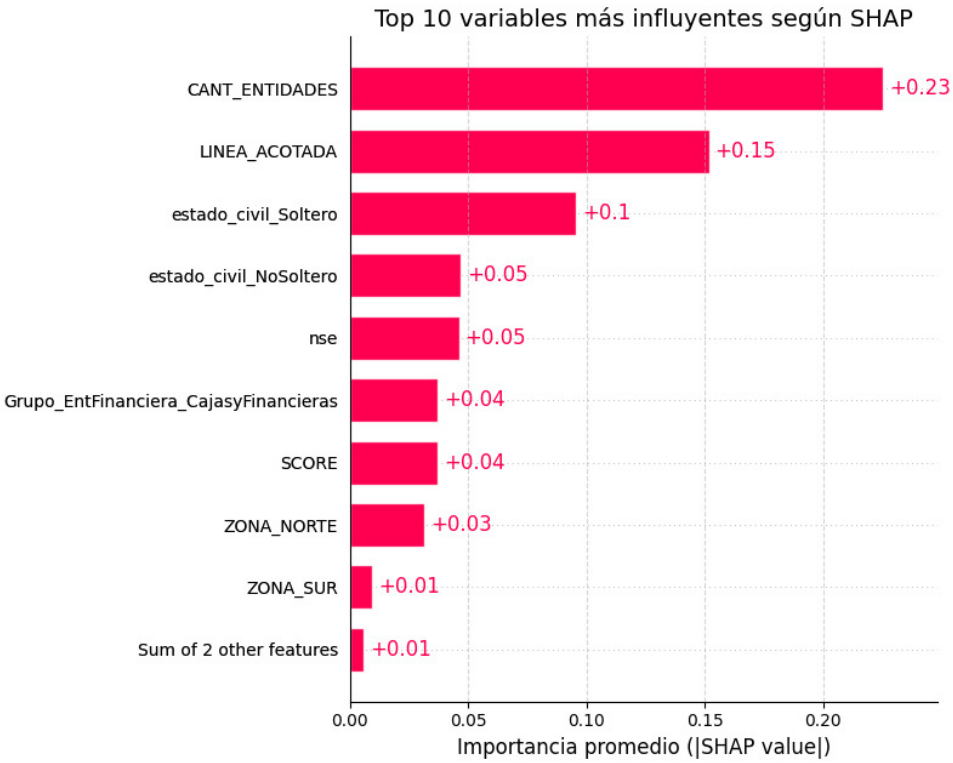
Pese a que los métodos coincidieron en gran parte, se decidió trabajar con el subconjunto generado por DecisionTree_Features, que seleccionó 9 variables finales. Luego del tratamiento con variables dummies y transformaciones necesarias, el modelo final utilizó 11 columnas para la predicción.

Para interpretar la importancia de cada variable, se utilizó el método SHAP, el cual permite entender el impacto promedio que cada variable tiene sobre las predicciones del modelo. La Figura 4 muestra las 10 variables más influyentes, destacando CANT_ENTIDADES, LINEA_ACOTADA y estado_civil_Soltero como las de mayor valor explicativo.

Este resultado evidencia que el modelo considera una combinación de variables financieras (LINEA_ACOTADA, SCORE), comportamiento histórico (CANT_ENTIDADES) y factores sociodemográficos (estado_civil, nse, zona) al momento de priorizar clientes. Así, se garantiza un enfoque multivariable y coherente con la lógica comercial del producto.

Figura 4

Importancia de variables en el modelo, valores medios absolutos de SHAP



Nota. Elaboración propia a partir de los valores promedio absolutos de SHAP generados con el modelo LDA final.

Durante la etapa de modelado, se evaluaron múltiples algoritmos de clasificación con el objetivo de identificar el que mejor equilibrara rendimiento predictivo, interpretabilidad y aplicabilidad operativa. Entre los modelos considerados se incluyeron árboles de decisión (Decision Tree, Random Forest), métodos de boosting (XGBoost, LightGBM), técnicas penalizadas (Lasso, Ridge) y redes neuronales básicas. El detalle de los algoritmos evaluados, junto con sus respectivas métricas, se encuentra en el Anexo 1: Comparativa de algoritmos candidatos.

Como parte del proceso de selección final, se identificaron dos modelos tentativos con mejor desempeño preliminar: RidgeClassifier y LinearDiscriminantAnalysis (LDA). A ambos se les aplicó un proceso de ajuste de hiperparámetros utilizando RandomGridSearch (ver

Figura 5), con el fin de optimizar su rendimiento. Los espacios de búsqueda definidos fueron los siguientes:

Figura 5

Espacios de búsqueda para RandomGridSearch



```
param_grids = {
    "Ridge Classifier": {
        "alpha": np.concatenate([np.logspace(-5, -1, 5), np.logspace(0, 3, 5)]),
        "solver": ["auto", "saga", "lsqr", "cholesky"],
        "fit_intercept": [True, False]
    },
    "LDA": {
        "solver": ["svd", "lsqr", "eigen"],
        "shrinkage": [None, "auto", 0.1, 0.5, 0.9]  # será ignorado si solver =
"svd"}
}
```

Nota. Fragmento de código elaborado por el autor para definir los espacios de búsqueda de hiperparámetros utilizados en el ajuste de modelos mediante RandomGridSearch.

Ambos modelos fueron entrenados bajo un esquema con datos balanceados utilizando **SMOTE (Synthetic Minority Over-sampling Technique)**, dada la naturaleza altamente desbalanceada del problema (tasa de aceptación < 1%). Además, se utilizó **MLflow** para realizar el seguimiento sistemático de todos los experimentos, permitiendo comparar de forma transparente las combinaciones evaluadas y sus métricas asociadas.

Finalmente, se optó por **LDA** como modelo definitivo, debido a que ofreció un mejor rendimiento en las métricas clave (mayor AUC y Recall) y además tiene la capacidad de retornar **probabilidades**, lo que facilita su integración en flujos comerciales que requieren priorización de clientes. El modelo logró un AUC de 0.713 (Gini = 0.426) en el conjunto de prueba, y evidenció ser una alternativa simple, robusta y trazable.

Con el objetivo de asegurar la reproducibilidad del modelo en ambientes productivos y facilitar su integración con flujos de datos reales, se construyó un **pipeline modular** que

contempla tanto el preprocesamiento de variables como la clasificación final mediante el modelo LDA previamente entrenado.

El pipeline de preprocesamiento se encarga de aplicar transformaciones categóricas, ordinales y numéricas según el tipo de variable, y fue persistido como archivo .pkl utilizando joblib. Del mismo modo, el modelo entrenado fue almacenado como objeto serializado y puede ser cargado directamente para realizar predicciones sobre nuevos datos.

La siguiente imagen (Fig. 6) muestra un ejemplo del proceso de carga y uso del pipeline para generar **scores de probabilidad**:

Figura 6

Ejemplo de código para cargar el pipeline y generar scores de probabilidad



```python
import joblib
from transformers.encoding import CategoricalOrdinalMapper
from transformers.utils import DataFrameWrapper
from transformers.data_validation import ValueFilterTransformer
from transformers.feature_engineering import CategoricalFeatureEngineer
import pandas as pd

from config.params_pipeline_process import *

data = pd.read_parquet(r'data/CLEAN/dataset_validate.parquet')

data = data[columnas_necesarias]
print(f"Shape: {data.shape}")


print("🚀 Cargando pipeline de preprocesamiento...", flush=True)
pipeline_preprocess = joblib.load(r'pipelines/pipeline_preprocesamiento.pkl')
pipeline_classifier = joblib.load(r'pipelines/model_clasification_lda.joblib')

expected_features = ['LINEA_ACOTADA', 'SCORE', 'CANT_ENTIDADES', 'RangoSaldo1', 'nse',
       'estado_civil_Soltero', 'Grupo_EntFinanciera_CajasyFinancieras',
       'PERFIL_VP2', 'ZONA_NORTE', 'ZONA_SUR', 'estado_civil_NoSoltero']

def predecir(data_validate, return_proba=False):
    X_validate = data_validate[columnas_necesarias]
    X_validate = pipeline_preprocess.transform(X_validate)

    if return_proba:
        y_proba = pipeline_classifier.predict(X_validate[expected_features])  # Esto ya retorna probas
en PyFuncModel
        return y_proba
    else:
        y_pred = pipeline_classifier.predict(X_validate[expected_features])
        return y_pred

data["predict_proba"] = predecir(data)

print("🚀 Generación de Scores de Probabilidad", flush=True)
print(data["predict_proba"].describe())
```

Tal como se observa, la función predecir() permite generar la probabilidad de aceptación de cada cliente (predict_proba), la cual es utilizada como score principal para la

priorización de contactos. Este enfoque facilita la aplicación del modelo en la operación diaria, permitiendo que los ejecutivos gestionen primero a los clientes con mayor probabilidad de conversión.

En resumen, el modelo desarrollado demostró ser útil y efectivo tanto en términos técnicos como operativos. Logró mejorar las tasas de aceptación y productividad posterior a su implementación, mantuvo un rendimiento estable en validación (AUC = 0.713, Gini = 0.426), y fue integrado mediante un pipeline robusto y trazable. Estas condiciones garantizan su aplicabilidad práctica en contextos reales y su potencial escalabilidad para futuras campañas comerciales.

IV. Conclusiones

El presente capítulo presenta las conclusiones obtenidas a partir del desarrollo e implementación del modelo de priorización de clientes en campañas comerciales. Estas conclusiones se encuentran alineadas con la finalidad general de la actividad, que fue optimizar la gestión comercial mediante el uso de técnicas de machine learning, y se organizan de acuerdo a los objetivos específicos planteados en la etapa inicial. Posteriormente, se detallan aspectos metodológicos, justificaciones técnicas y elementos operativos relevantes que complementan la evaluación integral del proyecto.

Cumplimiento de los objetivos planteados

En relación con la finalidad de la actividad —optimizar la gestión comercial mediante un sistema de priorización de clientes basado en datos—, se concluye que todos los objetivos fueron alcanzados de manera satisfactoria:

Exploración y análisis de datos históricos. Se consolidaron diversas fuentes internas, identificando patrones y variables relevantes mediante técnicas estadísticas y visuales.

Preparación del conjunto de datos. Se aplicaron procesos rigurosos de limpieza, imputación, codificación y transformación de variables, permitiendo estructurar un dataset robusto para tareas de clasificación.

Evaluación de modelos de clasificación. Se compararon más de diez algoritmos bajo un esquema de validación cruzada estratificada, utilizando métricas apropiadas para problemas desbalanceados como el AUC, Recall y F1-score.

Selección del modelo óptimo. Se seleccionó LDA tras un proceso de ajuste de hiperparámetros vía RandomGridSearch, destacando por su rendimiento (AUC = 0.713) y su capacidad de retorno de probabilidades.

Implementación del sistema de priorización. El modelo fue integrado en un pipeline reproducible y desplegado en producción. Los scores generados permitieron priorizar a los clientes más propensos, logrando mejoras sustanciales en la efectividad comercial.

Estos resultados demuestran que el enfoque analítico propuesto es viable, replicable y ha generado valor operativo tangible para la organización.

4.1. Justificación

La implementación de un modelo de machine learning para la priorización de clientes en campañas comerciales respondió a una necesidad concreta de optimización dentro del entorno operativo del call center. La estrategia previa, basada en juicio experto, no lograba capturar la complejidad de los patrones de comportamiento ni predecir con precisión la disposición de los clientes a aceptar un producto. Frente a ello, la propuesta desarrollada permitió establecer un enfoque más objetivo y eficaz, orientado por datos, que mejoró la productividad y efectividad de la campaña de manera tangible.

4.2. Metodología aplicada

Se siguió un enfoque estructurado basado en la metodología CRISP-DM, abarcando desde la comprensión del negocio hasta la generación de scores y su integración en la operación. Durante el proceso se aplicaron técnicas avanzadas de exploración y limpieza de datos, ingeniería de variables, análisis de multicolinealidad, y un proceso riguroso de

selección de características mediante métodos como Random Forest, RFE, Lasso, Boruta y otros.

Posteriormente, se evaluaron más de 10 algoritmos de clasificación, entre los que destacan LDA, RidgeClassifier, SVM, Random Forest y XGBoost. A partir de esta comparativa, se seleccionaron dos modelos tentativos —RidgeClassifier y LinearDiscriminantAnalysis— para someterlos a un ajuste de hiperparámetros mediante RandomGridSearch. Este proceso se realizó con datos balanceados utilizando SMOTE y fue monitoreado mediante MLflow, garantizando la trazabilidad y reproductibilidad de los experimentos.

El modelo LDA fue finalmente seleccionado por su rendimiento competitivo (AUC = 0.713, Gini = 0.426), su capacidad de retornar probabilidades y su bajo costo computacional. Además, su transparencia facilita la aceptación por parte del área de negocio. Como se señala en la literatura, los modelos lineales como LDA ofrecen una combinación adecuada de simplicidad, interpretabilidad y rendimiento en tareas de clasificación con datos estructurados (James, Witten, Hastie, & Tibshirani, 2013).

4.3. Descripción de la implementación

Para asegurar su aplicabilidad en entornos reales, se desarrolló una arquitectura modular basada en pipelines serializados con joblib, utilizando la librería scikit-learn como base para estructurar las etapas de preprocesamiento, transformación y clasificación del modelo (Scikit-learn Developers, 2024). La función predecir(), integrada al sistema, genera scores de probabilidad (predict_proba) que son utilizados para construir un ranking de clientes, priorizando a aquellos con mayor propensión a aceptar la oferta.

Este enfoque permitió traducir los resultados del modelo en decisiones operativas concretas, logrando que la gestión de clientes sea más eficiente y esté orientada por evidencia. A partir de su implementación en marzo de 2024, se observó un aumento sostenido en el monto desembolsado mensual, superando la meta establecida de S/ 1.2 millones. Asimismo, la efectividad entre clientes contactados (cet > 0) se duplicó en comparación con periodos anteriores.

En conjunto, el sistema desarrollado demostró ser técnicamente sólido, interpretable y útil en un entorno comercial real, generando valor cuantificable para la organización. Su diseño modular también garantiza su escalabilidad para futuras campañas u otros productos financieros.

V. Recomendaciones

Utilizar el ranking de clientes como herramienta de gestión diaria. Se recomienda que los supervisores y ejecutivos comerciales adopten el ranking de priorización generado por el modelo como base para definir el orden de contacto diario. Esto permitirá enfocar los primeros esfuerzos del día en los clientes con mayor probabilidad de aceptación, incrementando la tasa de conversión sin necesidad de aumentar el número de llamadas.

Asignar objetivos diferenciados según nivel de propensión. A partir del score de cada cliente, se sugiere agruparlos por niveles de propensión (alto, medio y bajo) y establecer metas comerciales diferenciadas para cada grupo. Esto permite una gestión más estratégica, evitando que el equipo dedique el mismo esfuerzo a clientes con bajo potencial de conversión.

Monitorear continuamente la efectividad por decil de score. Se recomienda que el área comercial realice un seguimiento mensual de la efectividad comercial por tramos de score. Esto permitirá validar el funcionamiento real del modelo y ajustar las estrategias de contacto si se detectan desviaciones o comportamientos inesperados en ciertos segmentos.

Capacitar al personal sobre el uso y valor del score. Para garantizar una adopción efectiva del modelo, es fundamental que tanto supervisores como ejecutivos comprendan qué representa el score, cómo se calcula y por qué es útil. Esto aumentará la confianza en el sistema y reducirá la resistencia al cambio en la forma tradicional de gestionar la cartera.

Incluir el score en los reportes operativos y tableros comerciales. Se recomienda incorporar el score de propensión como una columna adicional en los reportes y dashboards usados por el equipo comercial. Esto permitirá visualizar de forma integrada el potencial de

cada cliente junto a sus datos operativos, facilitando decisiones rápidas y alineadas a los objetivos de la campaña.

Redirigir esfuerzos comerciales hacia los clientes con baja efectividad histórica. Finalmente, se sugiere aprovechar los resultados del modelo para identificar segmentos con baja tasa de conversión recurrente. En lugar de insistir con múltiples contactos poco efectivos, estos clientes podrían recibir estrategias alternativas (como mensajes automatizados o campañas de refuerzo), liberando así al equipo comercial para enfocarse en perfiles más rentables.

VI. Bibliografía

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees. Monterey, CA: Wadsworth International Group. ISBN 978-0-412-04841-8*

Breiman, L. (2001). *Random forests. Machine Learning, 45(1), 5–32.* https://doi.org/10.1023/A:1010933404324

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide. The CRISP-DM Consortium.* https://www.the-modeling-agency.com/crisp-dm.pdf

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.* https://doi.org/10.1613/jair.953

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

Davenport, T. H., & Harris, J. G. (2007). *Competing on analytics: The new science of winning. Boston, MA: Harvard Business Press. ISBN 978-1-4221-0332-6*

Guyon, I., & Elisseeff, A. (2003). *An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.* https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Hand, D. J., & Henley, W. E. (1997). *Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523–541.* https://doi.org/10.1111/1467-985X.00082

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques (3rd ed.). Elsevier.* https://doi.org/10.1016/C2009-0-61819-5

Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers. ASQC Quality Press.* https://www.worldcat.org/title/27684661

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R. Springer.* https://www.statlearning.com

Kelleher, J. D., Mac Carthy, M., & Korvir, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies. MIT Press.* https://mitpress.mit.edu/9780262029445/fundamentals-of-machine-learning-for-predictive-data-analytics/

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling. Springer.* https://doi.org/10.1007/978-1-4614-6849-3

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data (2nd ed.). Wiley.* https://doi.org/10.1002/9781119013563

MLflow. (2018). *MLflow: Open source platform for the machine learning lifecycle.* https://mlflow.org

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking. O'Reilly Media.* https://www.oreilly.com/library/view/data-science-for/9781449374273/

Scikit-learn Developers. (2024). *scikit-learn: Machine learning in Python.* https://scikit-learn.org

Tukey, J. W. (1977). *Exploratory data analysis. Reading, MA: Addison-Wesley. ISBN 978-0-201-07616-5*

VII. ANEXOS e Ilustraciones

ANEXO I: Comparación de Algoritmos

Model	Accuracy	AUC	Recall	Prec.	F1	algoritmo_select_feature	Gini
SVM - Linear Kernel	0.675	0.7913	0.7784	0.0599	0.111	DecisionTree_modelos	0.5826
Ridge Classifier	0.7277	0.7976	0.7581	0.0683	0.1253	DecisionTree_modelos	0.5952
SVM - Linear Kernel	0.6829	0.7809	0.7541	0.059	0.1093	SelectFromModel_RF_modelos	0.5618
Linear Discriminant Analysis	0.7264	0.8058	0.7541	0.0677	0.1243	DecisionTree_modelos	0.6116
SVM - Linear Kernel	0.6989	0.7955	0.7541	0.0637	0.117	XGBoost_modelos	0.591
Quadratic Discriminant Analysis	0.7361	0.8019	0.7473	0.0696	0.1273	Random_Forest_modelos	0.6038
Linear Discriminant Analysis	0.7333	0.809	0.7419	0.0685	0.1253	all_features_modelos	0.618
Linear Discriminant Analysis	0.7333	0.809	0.7419	0.0685	0.1253	Lasso_Logic_modelos	0.618
Linear Discriminant Analysis	0.7333	0.809	0.7419	0.0685	0.1253	RFE_RF_modelos	0.618
Linear Discriminant Analysis	0.7306	0.8068	0.7405	0.0677	0.124	XGBoost_modelos	0.6136
Ridge Classifier	0.733	0.8015	0.7378	0.068	0.1246	XGBoost_modelos	0.603
Logistic Regression	0.7265	0.7935	0.7378	0.0665	0.1219	DecisionTree_modelos	0.587
Gradient Boosting Classifier	0.5328	0.6767	0.7365	0.0396	0.0751	LGBM_Metodo_modelos	0.3534
Ridge Classifier	0.7362	0.8046	0.7365	0.0688	0.1258	all_features_modelos	0.6092
Ridge Classifier	0.7362	0.8046	0.7365	0.0688	0.1258	RFE_RF_modelos	0.6092
Logistic Regression	0.7138	0.7841	0.7365	0.0635	0.1169	SelectFromModel_RF_modelos	0.5682
Ridge Classifier	0.7362	0.8046	0.7365	0.0688	0.1258	Lasso_Logic_modelos	0.6092
Ridge Classifier	0.7196	0.7899	0.7351	0.0647	0.1189	SelectFromModel_RF_modelos	0.5798
Naive Bayes	0.7368	0.7972	0.7338	0.0687	0.1256	SelectFromModel_RF_modelos	0.5944
SVM - Linear Kernel	0.7212	0.7983	0.7338	0.0663	0.1212	all_features_modelos	0.5966
SVM - Linear Kernel	0.7212	0.7983	0.7338	0.0663	0.1212	RFE_RF_modelos	0.5966
SVM - Linear Kernel	0.7212	0.7983	0.7338	0.0663	0.1212	Lasso_Logic_modelos	0.5966
Linear Discriminant Analysis	0.7279	0.8001	0.7297	0.0662	0.1213	Boruta_RF_modelos	0.6002

ANEXO II: Herramientas y tecnologías utilizadas

Herramienta / Tecnología	Descripción	Aplicación en el proyecto
Python 3.10	Lenguaje de programación de alto nivel	Desarrollo completo del pipeline, procesamiento de datos, entrenamiento y evaluación de modelos
pandas	Librería para análisis y manipulación de datos	Construcción del dataset, limpieza, transformación y agregación de información
scikit-learn	Librería de aprendizaje automático en Python	Implementación de modelos como LDA, RidgeClassifier, validación cruzada, métricas, preprocesamiento y pipelines
imblearn / SMOTE	Técnicas de muestreo para datos desbalanceados	Balanceo del conjunto de entrenamiento mediante sobre-muestreo sintético
joblib	Utilidad para la serialización de objetos en Python	Guardado y carga del pipeline y del modelo entrenado para producción
SHAP	Framework para la interpretabilidad de modelos de ML	Estimación del impacto de cada variable sobre las predicciones del modelo final
MLflow	Plataforma de seguimiento y gestión de experimentos de machine learning	Registro de experimentos, hiperparámetros y métricas durante la fase de tuning de modelos
matplotlib / seaborn	Librerías de visualización de datos	Generación de gráficos y análisis exploratorio de datos (EDA)
VS Code	Entorno de desarrollo integrado (IDE)	Desarrollo y ejecución del código fuente del proyecto
Jupyter Notebook	Entorno interactivo para Python	Prototipado, pruebas interactivas y visualización de resultados

ANEXO III: Ficha de Evaluación de TSP

FICHA DE EVALUACIÓN			
1) Datos Generales			
- Razón social de la empresa / institución	MF Asesoria y Consultoria SA		
- Área o Dpto. en que realizó la actividad	Inteligencia Comercial.		
- Nombre y cargo del profesional responsable	Luis Daniel Villacorta Tito		
- Periodo de labores (horas laboradas)	48 semanales.		
- Nombre del bachiller			
- EP			
- Facultad			
- Universidad			
2) Evaluación (*)	**Óptimo (2pts)**	**Bueno (1 pt)**	**Deficiente (0)**
a) Conocimientos	X		
b) Iniciativa (proactividad)	X		
c) Habilidad	X		
d) Cooperación (trabajo en equipo)	X		
e) Capacidad organizativa	X		
f) Eficiencia	X		
g) Dedicación	X		
h) Cumplimiento (trabajo bajo presión)		X	
i) Puntualidad		X	
j) Asistencia	X		
Sub total	16	2.	
TOTAL	18		

(*) Evaluación
Satisfactorio : 20 a 15
No satisfactorio : 14 a menos

