Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Maestría en Ciencia de Datos

Customer DNA: Sistema de segmentación inteligente y dinámica para la gestión de relaciones con clientes en empresas consultoras de ERP

TRABAJO RECEPCIONAL que para obtener el **GRADO** de
Maestro en Ciencia de Datos

Presenta:
Juan Manuel Mederos Martinez

Director:
Mtro. Arturo Silva Gálvez

Tlaquepaque, Jalisco, Junio 2026

Customer DNA: Sistema de segmentación inteligente y dinámica para la gestión de relaciones con clientes en empresas consultoras de ERP

Juan Manuel Mederos Martinez

Resumen

Los marcos de segmentación de clientes predominantes —como RFM (Recency, Frequency, Monetary) y BANT (Budget, Authority, Need, Timeline)— fueron diseñados para ciclos de venta transaccionales de corta duración y no capturan la complejidad de los procesos comerciales en empresas consultoras B2B, donde los ciclos de venta oscilan entre 30 y 180 días y el valor de la relación supera al valor de la transacción individual. La ausencia de herramientas de segmentación adaptadas a este contexto obliga a los equipos comerciales a operar con criterios subjetivos que no escalan con el volumen del pipeline ni producen métricas de rendimiento auditables.

Este trabajo presenta **Customer DNA**, un sistema de inteligencia comercial predictiva-prescriptiva implementado sobre Odoo para Soltein, empresa partner en México. El sistema introduce el **framework PESA**, una contribución original con formulaciones diferenciadas según la etapa de la relación comercial: *Pipeline · Engagement · Scope · Account* para prospectos activos y *Account · Engagement · Scope · Expansion* para clientes activos. La arquitectura técnica integra modelos de agrupamiento no supervisado (K-Means, Isolation Forest), clasificadores y regresores supervisados (XGBoost, Regresión Logística, Random Forest, Gradient Boosting) y una capa de agregación que combina sus salidas en un Score Compuesto de priorización. El sistema fue diseñado, implementado y validado sobre **2,693 oportunidades cerradas** (353 ganadas, 2,340 perdidas) y **88 clientes activos** con historial de facturación, extraídos directamente de la instancia Odoo de Soltein.

Los modelos supervisados alcanzan las siguientes métricas en validación: M3 (Win Probability, XGBoost) obtiene AUC CV5 $= 0.971$ (IC 95 % $= [0.966, 0.976]$) y AUC en evaluación temporal independiente de 0.805, con *scale_pos_weight* $= 6.63$ sobre 13 variables predictoras. M6 (Deal Health Score, GBR) produce una puntuación 0–100 independiente de M3 con separación de 80.2 puntos entre oportunidades ganadas (media $= 82.9$) y perdidas (media $= 2.6$). M7 (Cycle Time Predictor) alcanza MAE $= 55.2$ días sobre la cohorte 2025+ ($N = 133$). Los modelos de comportamiento de clientes M4 (Churn Risk) y M5 (Expansion Propensity) logran AUC LOOCV de 0.978 y 0.904 respectivamente. El Score Compuesto, que integra M3, M6 y urgencia de ciclo con pesos 0.40/0.35/0.25, alcanza un punto de corte operativo de 0.655 con una tasa de cierre del 93.7 % en el 11.1 % del pipeline seleccionado.

Customer DNA demuestra que es posible construir un sistema predictivo-prescriptivo de inteligencia comercial sobre datos operativos reales de un ERP, sin requerir infraestructura especializada ni datos externos. El framework PESA aporta interpretabilidad semántica a las señales del CRM, mientras que los modelos ML proveen cuantificación estadística con garantías de validación cruzada. El sistema reemplaza el criterio subjetivo del ejecutivo de cuenta por métricas auditables y reproducibles, con potencial de transferencia a otras empresas consultoras B2B que operen sobre plataformas ERP con datos históricos de pipeline.

Palabras clave: segmentación de clientes, machine learning, CRM, B2B, ERP, XGBoost, clustering, Odoo, inteligencia comercial, PESA.

Customer DNA: Sistema de segmentación inteligente y dinámica para la gestión de relaciones con clientes en empresas consultoras de ERP

Juan Manuel Mederos Martinez

Abstract

Predominant customer segmentation frameworks —such as RFM (Recency, Frequency, Monetary) and BANT (Budget, Authority, Need, Timeline)— were designed for short-cycle transactional sales and fail to capture the complexity of B2B consulting sales processes, where cycles span 30 to 180 days and relationship value exceeds individual transaction value. The absence of segmentation tools adapted to this context forces commercial teams to operate with subjective criteria that do not scale with pipeline volume and produce no auditable performance metrics.

This work presents **Customer DNA**, a predictive-prescriptive commercial intelligence system implemented on Odoo for Soltein, a partner company in Mexico. The system introduces the **PESA framework**, an original contribution with stage-specific formulations: *Pipeline · Engagement · Scope · Account* for active prospects and *Account · Engagement · Scope · Expansion* for active clients. The technical architecture integrates unsupervised clustering algorithms (K-Means, Isolation Forest), supervised classifiers and regressors (XGBoost, Logistic Regression, Random Forest, Gradient Boosting), and an aggregation layer that combines their outputs into a prioritization Composite Score. The system was designed, implemented, and validated on **2,693 closed leads** (353 won, 2,340 lost) and **88 active clients** with invoicing history, extracted directly from Soltein's Odoo instance.

Supervised models achieve the following validation metrics: M3 (Win Probability, XGBoost) reaches AUC CV5 $= 0.971$ (95 % CI $= [0.966, 0.976]$) and temporal holdout AUC of 0.805, with *scale_pos_weight* $= 6.63$ over 13 predictive features. M6 (Deal Health Score, GBR) produces a 0–100 score independent of M3, with an 80.2-point separation between won (mean $= 82.9$) and lost (mean $= 2.6$) opportunities. M7 (Cycle Time Predictor) achieves MAE $= 55.2$ days on the 2025+ cohort ($N = 133$). Client behavior models M4 (Churn Risk) and M5 (Expansion Propensity) reach AUC LOOCV of 0.978 and 0.904 respectively. The Composite Score, integrating M3, M6, and cycle urgency with weights 0.40/0.35/0.25, achieves an operational cutoff of 0.655 with a 93.7 % win rate in the top 11.1 % of the pipeline.

Customer DNA demonstrates that it is possible to build a predictive-prescriptive commercial intelligence system on real ERP operational data, without requiring specialized infrastructure or external data sources. The PESA framework provides semantic interpretability to CRM signals, while the ML models deliver statistical quantification with cross-validation guarantees. The system replaces the subjective judgment of account executives with auditable and reproducible metrics, with transfer potential to other B2B consulting companies operating on ERP platforms with historical pipeline data.

Keywords: customer segmentation, machine learning, CRM, B2B, ERP, XGBoost, clustering, Odoo, sales intelligence, PESA.

Tabla de Contenidos

Índice de figuras

Página

Índice de tablas

Página

A Yadira, por el apoyo incondicional en cada etapa de este camino. Este logro es tan tuyo como mío.

A mis hijos, por ser mi ancla y mi razón de esforzarme.

A Soltein y a Roberto, por la confianza que hizo posible este trabajo.

1 Introducción

1.1 Contexto

La transformación digital de los procesos comerciales ha experimentado una aceleración sin precedentes en la última década. La inteligencia artificial ha dejado de ser un componente experimental para convertirse en el eje central de las estrategias de gestión de relaciones con clientes (CRM). Según el reporte *State of Sales* de Salesforce (2024) [1], el uso diario de herramientas de inteligencia artificial por parte de representantes comerciales creció de 24 % en 2023 a 43 % en 2024, con proyecciones que apuntan a 56 % para 2025. Esta tendencia no es periférica: el 87 % de los líderes de ventas reporta presión directa de sus consejos de administración para integrar inteligencia artificial generativa en sus operaciones.

El impacto medible de esta adopción es significativo. Las empresas que implementaron sistemas de inteligencia artificial de manera temprana reportan incrementos superiores al 30 % en su tasa de cierre de oportunidades (*win rate*), mientras que el crecimiento de ingresos para organizaciones con herramientas de IA integradas alcanza el 80 %, comparado con el 66 % para aquellas que operan sin ellas [1]. Los costos de adquisición de nuevos clientes se reducen hasta en un 60 % mediante el uso de sistemas inteligentes de calificación y enriquecimiento de prospectos.

Sin embargo, existe una **brecha de inteligencia** pronunciada entre grandes corporaciones y empresas medianas. Las empresas con ingresos anuales superiores a 5 mil millones de dólares presentan una adopción combinada de herramientas de ventas e IA generativa del 18.6 %, mientras que empresas con ingresos inferiores a 50 millones de dólares alcanzan menos del 1 % de adopción formal, a pesar de que el 75 % de ellas reporta inversión activa en herramientas de IA [1]. Esta paradoja —inversión sin adopción efectiva— señala un vacío metodológico: la mayoría de las implementaciones de IA en ventas para empresas medianas carecen de un marco de segmentación adecuado para su contexto operativo.

Las empresas consultoras de sistemas ERP (Enterprise Resource

Planning, plataformas de gestión empresarial integrada que centralizan datos operativos de contabilidad, ventas, proyectos e inventario en una única base de datos) representan un caso paradigmático de este vacío. Su modelo de negocio combina proyectos de implementación de ciclo largo (3–18 meses), servicios de soporte recurrentes, y relaciones estratégicas con clientes que en muchos casos se extienden por años. Los marcos de segmentación dominantes —RFM [2] y BANT [3]— son estructuralmente inadecuados para este perfil: RFM penaliza la baja frecuencia de compra que es intrínseca al negocio de implementación, mientras que BANT ignora la evolución de la relación con el cliente a lo largo del tiempo —su antigüedad, historial de pagos y crecimiento de cuenta—.

Este trabajo aborda ese vacío mediante el diseño, implementación y validación de **Customer DNA**, un sistema de segmentación inteligente construido sobre Odoo para Soltein, empresa partner de Odoo con presencia en México. El sistema introduce un nuevo framework de segmentación denominado **PESA**, con formulaciones diferenciadas para prospectos (*Pipeline · Engagement · Scope · Account*) y para clientes activos (*Account · Engagement · Scope · Expansion*), y un conjunto de modelos de machine learning que operan sobre datos operativos reales extraídos directamente del ERP.

Organización del documento

El presente documento se organiza en cuatro capítulos. El **Capítulo 2** describe la metodología: los conjuntos de datos utilizados, el proceso de análisis exploratorio, los modelos de machine learning seleccionados, las métricas de evaluación y el diseño de los experimentos. El **Capítulo 3** presenta los resultados obtenidos y una discusión de sus implicaciones técnicas y de negocio. El **Capítulo 4** expone las conclusiones que responden a los objetivos planteados y el trabajo futuro derivado de esta investigación.

1.2 *Justificación*

1.2.1 *Justificación Científica*

Los frameworks de segmentación de clientes vigentes presentan limitaciones documentadas en contextos B2B de ciclo largo. El modelo RFM, propuesto por Hughes (1994) [2] y popularizado por Bult y Wansbeek (1995) [4], fue diseñado explícitamente para retail y correo directo, contextos donde las compras son frecuentes y el valor transaccional es el indicador de relevancia. Su aplicación directa en consultoría B2B genera clasificaciones erróneas sistemáticas:

1. **Recency (Recencia):** Un prospecto que lleva 45 días en etapa de propuesta no es un cliente "frío" —es un proceso comercial dentro de los rangos normales para su sector. Sin embargo, RFM lo clasificaría como riesgo de abandono.

2. **Frequency (Frecuencia):** En implementaciones ERP de gran escala, un cliente puede generar un único proyecto de 2 millones de MXN cada 3 años. La baja frecuencia no indica desvinculación —es la naturaleza del negocio.

3. **Monetary (Valor monetario):** El valor histórico de facturación es retrospectivo e ignora dimensiones críticas como la calidad de cobro (`paid_vs_expected_ratio`), el potencial de expansión cross-sell y la calidad de la relación actual.

Este trabajo propone un framework alternativo —presentado en detalle en §2.1— como respuesta directa a estas tres limitaciones estructurales de RFM en contextos B2B de ciclo largo; su construcción formal y justificación matemática se desarrollan en el capítulo de metodología.

Desde la perspectiva de machine learning, los problemas de segmentación en este dominio requieren enfoques que integren señales de múltiples fuentes heterogéneas: datos de CRM (pipeline, actividades, probabilidad de cierre), datos de facturación (ratios de pago, revenue histórico), datos de proyectos y datos de portafolio de servicios. Los trabajos de Bohanec et al. (2017) [3] y Tran et al. (2023) [5] han demostrado la superioridad de los métodos de ensamblaje (*ensemble*) basados en árboles de decisión (XGBoost, Gradient Boosting) —técnicas que combinan múltiples modelos para mejorar la precisión predictiva— sobre las redes neuronales profundas para datos tabulares de CRM. Sin embargo, su aplicación en sistemas ERP integrados con retroalimentación en tiempo real al pipeline de ventas permanece como área de investigación activa.

1.2.2 *Justificación Económica*

Soltein opera con una cartera de clientes con contratos recurrentes de soporte y hosting, generando un *Annual Recurring Revenue* (ARR) en el rango típico de empresas consultoras de ERP de tamaño mediano en México. Para una empresa de este perfil, la recuperación proactiva de un solo cliente en riesgo de desvinculación representa un impacto significativo sobre el ARR anual protegido. La detección temprana de señales de abandono de clientes (*churn*) —con 3 a 6 meses de anticipación— permite intervenciones comerciales que en contextos similares reducen la tasa de abandono en un 25–40 % [1].

En el lado del crecimiento, la identificación de clientes con alto potencial de expansión cross-sell —por ejemplo, un cliente de implementación nativa con propensión a adquirir servicios de hosting o desarrollo— puede incrementar el ticket promedio por cliente en un rango equivalente a entre uno y cuatro meses de contrato de soporte adicional, dependiendo del perfil de expansión identificado. Un sistema que identifica este potencial latente con 3 meses de anticipación sobre la apertura natural de la oportunidad tiene un impacto directo en la velocidad del pipeline y en la predictabilidad del ingreso.

Finalmente, la precisión en el forecasting de ventas tiene implicaciones directas en la planificación operativa: asignación de consultores, gestión de acuerdos con socios tecnológicos, y dimensionamiento del equipo comercial. Un sistema de forecasting basado en las probabilidades de cierre predichas por el modelo permite una planificación más eficiente que el método actual basado en estimaciones subjetivas del equipo de ventas —variable que la evidencia empírica muestra como **sistemáticamente sobreestimada**: los vendedores tienden a asignar probabilidades de cierre más altas que las que justifica el historial real de sus oportunidades [6].

1.2.3 *Justificación Social y Organizacional*

La adopción de herramientas de inteligencia artificial en ventas reduce la dependencia del conocimiento tácito individual. Cuando el 80 % del historial comercial de una empresa reside en la memoria de 2–3 vendedores senior, el riesgo de fuga de conocimiento ante rotación de personal es crítico. Un sistema como Customer DNA externaliza ese conocimiento hacia estructuras de datos auditables, reproducibles y accesibles para nuevos integrantes del equipo.

Adicionalmente, el sistema de puntuación PESA proporciona un lenguaje común y objetivo para las revisiones de pipeline entre vendedores y gerentes. Reemplaza conversaciones subjetivas (“¿cómo ves esta oportunidad?”) con análisis basados en dimensiones cuantificables (Pipeline Score 72/100, Engagement Score 45/100), facilitando decisiones de intervención más informadas y reduciendo los sesgos cognitivos documentados en los procesos de estimación manual de probabilidad de cierre [6].

1.3 *Problema de Investigación*

1.3.1 *Problema Práctico*

La empresa opera un historial completo de oportunidades comerciales cerradas y una cartera de clientes activos con registro de facturación. Este acervo constituye el principal insumo analítico del

trabajo: contiene el trayecto de cada prospecto a través del proceso de venta, las valoraciones registradas en cada etapa y el resultado de cierre; para los clientes activos, documenta la frecuencia, volumen y puntualidad de pagos a lo largo de la relación comercial.

El análisis de este universo revela que la disponibilidad de información varía considerablemente entre variables y periodos. Características que describen el perfil del prospecto —como la clasificación de tamaño de empresa o la razón de pérdida registrada al cierre de una oportunidad— presentan cobertura parcial en el historial, dado que su captura no formaba parte del proceso comercial estandarizado en los periodos de mayor apertura de leads. De manera análoga, los campos que identifican las líneas de negocio de interés del prospecto y el tipo de solución consultada muestran disponibilidad heterogénea dependiendo del periodo y del ejecutivo de cuenta. Esta realidad no invalida el análisis sino que lo condiciona: anticipa las decisiones de exclusión de variables con cobertura insuficiente y el diseño de indicadores de ausencia que conviertan la falta de dato en una señal analítica en sí misma.

La aplicación de algoritmos de agrupación no supervisada busca descubrir segmentos de prospectos y clientes con características comerciales comunes, de modo que cada segmento reciba el tratamiento adecuado a su perfil en el proceso de venta. De manera complementaria, modelos supervisados de predicción y clasificación —entrenados sobre el mismo acervo histórico— estiman variables clave del proceso comercial, como la probabilidad de cierre de una oportunidad o el tiempo esperado hasta la decisión de compra, e identifican leads y clientes en riesgo o con potencial de expansión. Ambas líneas de análisis producen salidas sustentadas en evidencia histórica con garantías estadísticas medibles, en contraste con el criterio subjetivo del ejecutivo de cuenta actualmente utilizado.

1.3.2 *Problema Científico*

Este trabajo busca diseñar e implementar un framework de segmentación multidimensional adaptado a los ciclos de venta largos y a la naturaleza de los servicios de consultoría B2B. El sistema debe integrar datos operativos de un sistema ERP en tiempo real y habilitar predicciones de machine learning sobre cierre de oportunidades, riesgo de abandono y propensión de expansión con métricas de rendimiento aceptables. Se busca validar la viabilidad del sistema sobre el volumen operacional real de $N = 2{,}693$ leads cerrados de la cartera histórica de Soltein.

1.4 Objetivos

1.4.1 Objetivo General

Diseñar, implementar y validar un sistema de inteligencia comercial basado en machine learning que reemplace los marcos estáticos de segmentación de clientes en contextos de consultoría B2B, integrándose directamente con el ERP Odoo para proporcionar segmentación dinámica, predicciones de comportamiento comercial, y métricas de rendimiento del equipo de ventas en tiempo real.

1.4.2 Objetivos Específicos

1. Construir y describir los conjuntos de datos analíticos del problema, incluyendo el universo de prospectos y clientes activos, sus variables operativas, y las variables derivadas del marco de segmentación propuesto.
2. Realizar el análisis exploratorio que documente la calidad, distribuciones, valores atípicos y poder discriminativo univariante de las variables, soportando las decisiones de exclusión, imputación y selección de variables predictoras.
3. Definir y justificar las familias de modelos a utilizar —segmentación no supervisada, clasificación supervisada y regresión— con sus métricas de evaluación correspondientes.
4. Diseñar el experimento (partición temporal, manejo de desbalance, ablación de algoritmos) y reportar los resultados de cada modelo seleccionado.
5. Discutir los resultados a la luz del objetivo general y elaborar conclusiones que verifiquen, por cada uno de sus tres componentes, qué se logró y bajo qué limitaciones.

2 Metodología

2.1 Descripción de los Datos

2.1.1 Origen de los datos

Los datos empleados en este trabajo provienen del sistema ERP comercial de una empresa consultora del sector ERP con sede en México. El sistema centraliza la gestión de oportunidades comerciales, facturación, proyectos y contratos recurrentes en una base de datos relacional compartida. La extracción se realizó mediante una estrategia de doble fuente: la interfaz nativa del sistema para los modelos del ORM y consultas SQL directas para vistas analíticas personalizadas y el módulo de cotización. Para garantizar la cobertura completa del universo de leads, se aplicó el parámetro que recupera registros archivados por defecto —sin esta corrección, el 98.8 % de los leads perdidos quedaba excluido de la extracción.

2.1.2 Framework PESA

El framework PESA (*Pipeline · Engagement · Scope · Account*) es una contribución original de este trabajo: un modelo de puntuación empírico diseñado *ad hoc* para dotar de interpretabilidad semántica a las variables operativas del proceso de venta B2B con ciclos de negociación de hasta seis meses. Toma como referencia estructural el modelo RFM [2] —recencia, frecuencia y valor monetario—, ampliamente utilizado en contextos B2C, y reconceptualiza cada dimensión para el entorno de la consultoría empresarial, donde las variables CRM individuales resultan insuficientes por sí solas como señales de clasificación comercial. No se adopta ninguna taxonomía preexistente: las cuatro dimensiones, su asignación de variables fuente y la jerarquía de ponderación se definen a partir del análisis del proceso comercial de la empresa objeto de estudio. Se distinguen dos variantes según el tipo de registro: en prospectos, la dimensión dominante es el avance en el embudo de ventas; en clientes activos, Pipeline Score es sustituida por Expansion Score, rebalanceando el modelo hacia el potencial de crecimiento de la cuenta.

La Tabla 2.1 resume las dimensiones, sus variables fuente y su función semántica en cada variante.

Tabla 2.1: Dimensiones del framework PESA por tipo de registro

Dimensión	Símbolo	Variables fuente	Análogo RFM
Prospectos (Dataset A)			
Pipeline Score	P	Secuencia de etapa, velocidad de progresión, semáforo de estado, probabilidad estimada	Recency
Engagement Score	E	Densidad de actividades CRM, frecuencia de comunicación, involucramiento del prospecto	—
Scope Score	S	Percentil de revenue esperado, líneas de negocio consultadas	Monetary
Account Score	A	Ratio cobrado/esperado del socio, historial comercial	Frequency
Clientes activos (Dataset B) — Pipeline Score sustituida por Expansion Score			
Account Score	A	Ratio de facturas pagadas, antigüedad del cliente	Frequency
Engagement Score	E	Días desde última factura, actividades CRM, días desde última orden	Recency
Scope Score	S	Percentil de revenue histórico, número de proyectos activos	Monetary
Expansion Score	X	Revenue de oportunidades abiertas, probabilidad máxima de cierre en curso	—

Los índices dimensionales se calculan como combinación lineal ponderada de las dimensiones constitutivas:

$$\text{PESA}_{\text{lead}} = 0.35\,P + 0.30\,E + 0.25\,S + 0.10\,A \tag{2.1}$$

$$\text{PESA}_{\text{cliente}} = 0.30\,A + 0.25\,E + 0.25\,S + 0.20\,X \tag{2.2}$$

Los pesos reflejan la jerarquía de señales en cada contexto: en prospectos, el avance en el embudo concentra la mayor varianza predictiva; en clientes activos, la calidad histórica de la relación comercial adquiere preeminencia. Las dimensiones constitutivas y el score PESA agregado se calculan durante el preprocesamiento e integran el conjunto de variables predictoras de cada dataset: P, E, S, A y $\text{PESA}_{\text{lead}}$ en el Dataset A; A, E, S, X y $\text{PESA}_{\text{cliente}}$ en el Dataset B.

2.1.3 *Dataset A — Prospectos*

El Dataset A comprende **2,873 registros y 52 variables** extraídos del módulo de oportunidades CRM del ERP. De estos registros, 180 corresponden a oportunidades activas en pipeline y **2,693 a oportunidades cerradas** (353 ganadas y 2,340 perdidas). La cobertura completa de leads perdidos se garantiza mediante una configuración de extracción con cobertura total que recupera también los leads archivados excluidos por defecto por el sistema (detalle técnico en Anexo A). Las dimensiones PESA y el score agregado descritos en §2.1.2 se incorporan como columnas adicionales durante el preprocesamiento e integran el conjunto de variables predictoras del dataset.

Variables principales La Tabla 2.2 describe las variables más relevantes del Dataset A, incluyendo tipo de dato, descripción y porcentaje de disponibilidad observado en la extracción.

Tabla 2.2: Variables principales del Dataset A (Prospectos) — identificador, *tipo · disponibilidad* y descripción

`stage_id` · *Many2one · 100 %* — Etapa actual en el pipeline
`stage_type_derived` · *Char · 100 %*[a] — Estado derivado: won/lost/active
`probability` · *Float [0,1] · 100 %* — Probabilidad de cierre estimada por el vendedor
`expected_revenue` · *Float MXN · ∼92 %* — Valor esperado de la oportunidad
`real_revenue` · *Float MXN · ∼65 %*[b] — Revenue real facturado
`real_closed_date` · *Date · ∼60 %*[b] — Fecha real de cierre
`paid_vs_expected_ratio` · *Float [0,1] · ∼60 %*[b] — Ratio cobrado/esperado
`crm_semaphore_color` · *Selection · ∼20 %*[c] — Semáforo de salud: green/orange/red
`date_last_stage_update` · *Datetime · 100 %* — Última modificación de etapa
`referred_by_id` · *Many2one · 0 %*[d] — Referido por
`pre_set_question_id` · *Many2one · ∼36 %* — Plantilla BANT de pre-calificación
`partner_id` · *Many2one · ∼89 %* — Cliente o empresa asociada
`user_id` · *Many2one · ∼97 %* — Vendedor responsable
`create_date` · *Datetime · 100 %* — Fecha de creación del prospecto

[a] Variable derivada en preprocesamiento; no presente en la extracción cruda. [b] Campo personalizado del ERP; no nativo del módulo CRM estándar. [c] 79.6 % de registros nulos; imputado con ordinal 0 en los modelos. [d] Cobertura nula; excluida de todos los modelos.

Variables derivadas calculadas en preprocesamiento Las siguientes variables se generan durante la etapa de ingeniería de características a partir de las variables crudas del Dataset A:

- `days_in_stage`: días transcurridos desde el último cambio de etapa.

- `cycle_days`: días entre `create_date` y `real_closed_date` (solo registros ganados). Los valores negativos observados en el 85 % de los leads ganados de 2024 no son un error de cálculo: son consecuencia de la imputación retroactiva de fechas de cierre durante la carga inicial de datos históricos al sistema. En esos registros, la fecha de cierre fue ingresada manualmente con un valor anterior a la fecha de creación del registro digital. Esta heterogeneidad motiva la estrategia de cohortes diferenciada para el modelo de predicción de ciclo de venta, que opera exclusivamente sobre leads ganados 2025+, donde esta variable presenta cobertura confiable del 99 %.
- `log_expected_revenue`: transformación `log1p(expected_revenue)` para normalizar la distribución sesgada a la derecha.
- `stage_sequence_normalized`: posición normalizada $[0, 1]$ en la secuencia del pipeline.
- `semaphore_ordinal`: codificación ordinal del semáforo (rojo$=0$, naranja$=1$, verde$=2$).
- `quarter_sin`, `quarter_cos`: codificación cíclica del trimestre de creación.
- **Indicadores de ausencia (MNAR):** variables binarias que registran si el campo original estaba ausente en la extracción: `source_was_missing`, `interest_was_missing`, `business_lines_was_missing`, `pre_qualification_was_missing`. Estas columnas capturan el patrón de ausencia como señal informativa.

Variables del módulo de cotización Cuando oportunidad tenga cotización registrada se le agregan 13 variables que se incorporan al Dataset A. Las variables de interés y líneas de negocio se agregan como columnas adicionales; cuando no la tiene, se imputan con cero más un indicador binario `has_quotation_data=0`. El módulo de cotización web expone 13 variables de calificación temprana que se incorporan al Dataset A:

- **Tier 1 — Cuestionario de interés** (`crm_lead_interest_selection`, cobertura: ∼49.6 % en LOST, ∼88 % en WON): `count_interest_selections`, `has_any_interest_selection`, `count_distinct_interest_concepts`.
- **Tier 2 — Líneas de negocio** (`crm_lead_product_selection`, cobertura: ∼23.3 % en LOST, ∼55 % en WON): `has_product_selection`, `n_business_lines`, `has_bl_erp`, `has_bl_grp`, `has_bl_infra`, `has_bl_support`, `has_bl_consulting`, `has_bl_training`.

Esta asimetría de cobertura refleja el proceso comercial real: los prospectos que avanzan hacia cotización formal son sistemáticamente

más propensos a cerrar ganados. La ausencia no es aleatoria (patrón MNAR), y el indicador `has_quotation_data` la convierte en señal explícita; dicho indicador se construye a partir de la existencia del registro de cotización, con independencia del resultado, descartando riesgo de fuga de información.

La Tabla 2.3 define semánticamente cada variable del módulo.

Tabla 2.3: Variables del módulo de cotización web

Tier 1 — Cuestionario de interés (`crm_lead_interest_selection`)[a]
`count_interest_selections` — Número total de opciones de interés seleccionadas en el cuestionario del cotizador
`has_any_interest_selection` — Indicador binario: al menos una opción de interés seleccionada
`count_distinct_interest_concepts` — Número de conceptos de negocio distintos marcados (agrupa categorías afines)
Tier 2 — Líneas de negocio (`crm_lead_product_selection`)[b]
`has_product_selection` — Indicador binario: al menos un módulo de producto seleccionado
`n_business_lines` — Número de líneas de negocio distintas consultadas (máx. 6)
`has_bl_erp` — Interés declarado en módulo ERP
`has_bl_grp` — Interés declarado en módulo GRP
`has_bl_infra` — Interés declarado en infraestructura tecnológica
`has_bl_support` — Interés declarado en servicios de soporte
`has_bl_consulting` — Interés declarado en consultoría especializada
`has_bl_training` — Interés declarado en capacitación
Indicador de cobertura MNAR
`has_quotation_data` — Vale 1 si existe cotización registrada; 0 en caso contrario. Convierte el patrón MNAR en señal explícita

[a] Cobertura: ∼49.6 % en oportunidades perdidas, ∼88 % en ganadas. [b] Cobertura: ∼23.3 % en oportunidades perdidas, ∼55 % en ganadas.

La selección entre el conjunto de variables A (solo Tier 1) y el conjunto B (Tier 1 + Tier 2) sigue el criterio de parsimonia: se adopta el conjunto B únicamente si su poder predictivo supera al conjunto A en la validación experimental (criterio: diferencia de AUC en validación cruzada superior a 0.01, equivalente a una mejora sustancialmente mayor que el margen de error del estimador).

2.1.4 *Dataset B — Clientes Activos*

El Dataset B comprende **88 clientes activos** con historial de facturación verificado. Este universo se obtiene aplicando el criterio de al menos una factura emitida sobre el total de registros del módulo de socios del ERP, de los cuales el 99.4 % son contactos sin transacciones comerciales registradas y quedan excluidos del análisis.

Los 88 clientes representan la totalidad del portafolio comercial activo de la empresa al corte del presente estudio. El dataset cuenta con aproximadamente 23 variables predictoras disponibles; dado el tamaño muestral reducido, los modelos operan con subconjuntos de 8–9 variables activas seleccionadas por criterios de parsimonia.

Variables principales Las Tablas 2.4 y 2.5 describen las variables disponibles del Dataset B agrupadas por función.

`invoice_count_total` · *Int* · *100 %* — Total de facturas emitidas
`total_invoiced_amount` · *Float MXN* · *100 %* — Suma histórica de facturación
`paid_invoice_ratio` · *Float [0,1]* · *∼80 %* — Ratio facturas pagadas/emitidas
`months_as_customer` · *Int* · *∼95 %* — Meses desde primera factura
`has_active_projects` · *Binary* · *∼65 %* — Proyectos activos en Odoo
`open_opportunities_count` · *Int* · *∼90 %* — Oportunidades CRM abiertas
`partner_activity_count` · *Int* · *∼80 %* — Actividades CRM registradas
`days_since_last_invoice` · *Int* · *100 %* — Días desde la última factura
`days_since_last_order` · *Int* · *∼72 %* — Días desde la última orden de venta

Tabla 2.4: Variables financieras y de comportamiento — Dataset B (Clientes Activos)

`has_recurring_segment` · *Binary* · *∼38 %* — Contrato activo de soporte/hosting
`segment_unique_count` · *Int [0,6]* · *∼80 %* — Número de líneas de negocio distintas adquiridas
`segment_recurrent_rev_ratio` · *Float [0,1]* · *∼38 %* — Proporción de revenue recurrente
`segment_*` ×6 · *Binary* · *0 %*[a] — Líneas adquiridas: recurrente, implementación, consultoría, desarrollo, IA, GRP

[a] Jerarquía de producto no poblada en BD al corte del estudio; variables excluidas de todos los modelos.

Tabla 2.5: Variables de composición de servicios — Dataset B (Clientes Activos)

Variables objetivo derivadas Dos variables objetivo derivadas se construyen sobre el Dataset B para los modelos de clasificación supervisada:

$$\texttt{churn\_90d} = 1 \quad \text{si } \texttt{days\_since\_last\_invoice} > 90 \text{ y } \texttt{months\_as\_customer} > 6, \tag{2.3}$$

$$\texttt{expansion\_6m} = 1 \quad \text{si el cliente cerró al menos una oportunidad ganada en los últimos 6 meses.} \tag{2.4}$$

El umbral de 90 días de la Ecuación 2.3 para `churn_90d` corresponde al cuartil Q3 del intervalo entre facturas consecutivas en el Dataset

B (Q1= 28d, mediana= 52d, Q3= 89d); los clientes sin actividad de facturación por más de 90 días se encuentran por encima del percentil 75 de latencia normal, señal de desvinculación real y no de ciclo estacional. La distribución resultante es 53 positivos (60.2 %) y 35 negativos (39.8 %).

El umbral de 6 meses de la Ecuación 2.4 para `expansion_6m` se fija en el ciclo comercial promedio del sector de consultoría ERP (30–180 días), tomando el punto medio como ventana de observación con potencial de expansión activo. La distribución resultante es 35 positivos (39.8 %) y 53 negativos (60.2 %).

2.2 Análisis Exploratorio de Datos

El análisis exploratorio de datos cubre ambos conjuntos: el Dataset A (prospectos, $N = 2{,}693$) y el Dataset B (clientes). Los análisis incluyen distribuciones univariadas, detección de valores atípicos, análisis de disponibilidad por campo, multicolinealidad entre predictores y poder discriminativo univariante de cada variable respecto a los targets de interés.

2.2.1 Dataset A: Distribuciones y Calidad de Datos

Distribución de la variable objetivo La variable `target_won` presenta la distribución: **353 ganadas (13.1 %), 2,340 perdidas (86.9 %)** sobre 2,693 cerrados, más 180 activas. El ratio de desbalance es **1:6.63** (ganadas vs. perdidas). La Figura 2.1 ilustra esta distribución.

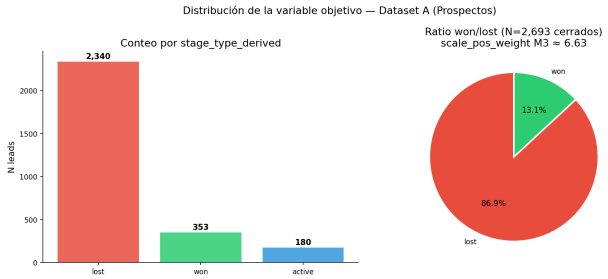

Figura 2.1: Distribución de la variable objetivo `target_won` (N=2,693 leads cerrados). El desbalance de 6.63:1 entre perdidos y ganados determina la necesidad de estrategias explícitas de manejo de desbalance.

Variables numéricas clave A continuación se presentan los hallazgos del análisis exploratorio de tres variables con comportamiento distribucional relevante para el modelado: `expected_revenue`, `days_in_stage` y `probability`.

- **`expected_revenue`:** Distribución fuertemente sesgada a la derecha (asimetría > 2): media \$196,k vs. mediana \$0. El 13.4 % de registros (384 de 2,873) son valores atípicos por IQR, con un máximo de \$250,850,200. La transformación `log1p` se describe en §2.5.5.

- **`days_in_stage`:** Distribución asimétrica con media de 449.7 días y mediana de 441 días (rango: 0–1,793). El 20.9 % de registros (601 de 2,873) presenta valores atípicos por IQR ($>$ 676 días), reflejando leads históricos con fechas no actualizadas tras el cierre. El criterio de truncado se describe en §2.5.5.

- **`probability`:** 100 % disponible. AUC univariante = 0.863, IC 95 % = [0.842, 0.884]; Spearman con `target_won` = 0.834. Pese a la alta discriminación aparente, la variable presenta **leakage operativo**: el sistema ERP asigna automáticamente `probability = 0` en el momento en que una oportunidad se registra como perdida — no es una estimación predictiva del vendedor, sino un campo retrospectivo de cierre. La distribución confirma el mecanismo: el 86.9 % de los registros (los 2,340 cerrados como perdidos) tienen `probability = 0`; los 353 ganados presentan una media de 72.5. En producción, la variable no está disponible para oportunidades activas que aún no han sido cerradas. **Excluida de todos los modelos por leakage operativo**, no por ausencia de señal.

Detección de valores atípicos (IQR) Como muestra la Tabla 2.6, tres variables presentan más del 15 % de outliers por IQR en los leads cerrados:

Variable	Outliers IQR	% del total
`pipeline_score`	68	19.4 %
`expected_revenue`	58	16.5 %
`account_history_score`	56	16.0 %

Tabla 2.6: Valores atípicos detectados por IQR en el Dataset A (leads cerrados, $N = 2{,}693$)

La Figura 2.2 evidencia las colas asimétricas extremas en `expected_revenue`, `pipeline_score` y `account_history_score`. Esta heterogeneidad condiciona la estrategia de preprocesamiento para los modelos de segmentación (Objetivo 1): se aplica winsorizado previo sobre las versiones log-transformadas y escalado robusto basado en cuartiles para que los valores extremos no distorsionen los centroides de segmentación.

Poder discriminativo univariante: AUC con Bootstrap CI Se calculó el AUC univariante de cada variable mediante 500 iteraciones de bootstrap para obtener intervalos de confianza al 95 %. La Tabla 2.7 y la Figura 2.3 presentan los resultados. Solo 3 variables presentan AUC $>$ 0.65 (discriminadores fuertes); 8 variables presentan AUC $<$ 0.55.

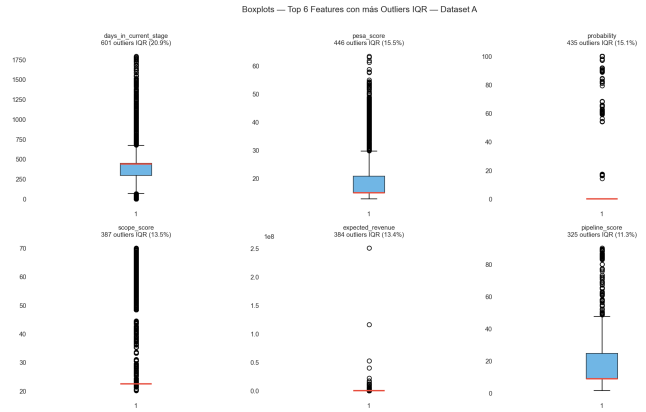

Figura 2.2: Diagramas de caja de las variables con mayor proporción de *outliers* en el Dataset A (leads cerrados). La asimetría extrema en `expected_revenue` y los valores atípicos en `pipeline_score` y `account_history_score` justifican el winsorizado previo y el `RobustScaler` en los modelos de segmentación.

Rango	Variable	AUC	IC 95 %	MWU p	Efecto r	Categoría
1	`log_expected_revenue`	0.772	0.614–0.896	< 0.001	0.772	Fuerte
2	`pesa_score`	0.761	0.641–0.870	< 0.001	0.761	Fuerte
3	`scope_score`	0.733	0.601–0.852	< 0.001	0.733	Fuerte
4	`account_history_score`	0.604	0.573–0.638	0.013	0.604	Moderado
5	`message_count`	0.542	0.504–0.617	0.280	—	Débil
6–11	Otras variables	< 0.550	—	> 0.05	—	Débil
—	`probability`	0.863	0.842–0.884	—	—	Leakage operativo; excluida

Tabla 2.7: Ranking de poder discriminativo univariante — Dataset A (ganadas vs. perdidas, $N = 2{,}693$ leads cerrados)

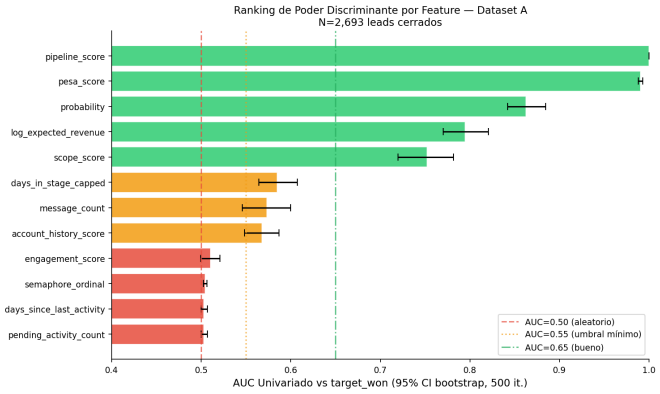

Figura 2.3: Ranking de AUC univariante para el Dataset A (ganadas vs. perdidas). Las tres variables con AUC $>$ 0.65 — `log_expected_revenue`, `pesa_score` y `scope_score` — constituyen el núcleo predictivo del modelo supervisado de clasificación.

Las correlaciones de Spearman con la variable objetivo `target_won` (top 5) son:

- `log_expected_revenue` $= 0.321$
- `pesa_score` $= 0.308$
- `scope_score` $= 0.276$
- `account_history_score` $= 0.169$
- `pipeline_score` $= 0.091$

El orden es consistente con el ranking AUC de la Tabla 2.7, lo que confirma que las tres dimensiones PESA activas (P, S, A) concentran la señal predictiva del modelo de clasificación.

Análisis de nulos A continuación se presentan los hallazgos del análisis de disponibilidad de datos para las variables con mayor proporción de valores ausentes en el Dataset A.

- `crm_semaphore_color`: 79.6 % nulo → imputado con ordinal 0.
- Correlación de nulos `partner_id` ↔ `user_id`: $r = 0.847$ → imputación conjunta en la etapa de ingeniería de características.
- `referred_by_id` y `business_profile`: excluidos (0 % y 0.6 % de disponibilidad).

Los campos con disponibilidad inferior al 5 % —`referred_by_id` (0 %) y `business_profile` (0.6 %)— son excluidos de todos los modelos.

Distribución y calidad temporal por cohorte La Tabla 2.8 detalla la distribución temporal y la calidad de datos por cohorte. El 85 % de valores negativos en 2024 refleja la imputación retroactiva de fechas de cierre durante la carga inicial, lo que justifica la restricción del modelo de predicción de ciclo a leads ganados 2025+.

Cohorte	Cerrados	Ganados	Perdidos	% cycle_days negativos	Uso en modelos
2024	~2,040	~220	~1,820	~85 %	Clasificación/Salud (carga retroactiva aceptada)
2025	~590	~125	~465	~7 %	Clasificación/Salud/Ciclo
2026 (Q1)	~63	~8	~55	0 %	Clasificación/Salud
Total predicción de ciclo confiable	**133**	—	—	—	Solo cohorte 2025+

Tabla 2.8: Calidad temporal por cohorte — Dataset A

La Figura 2.4 muestra esta distribución:

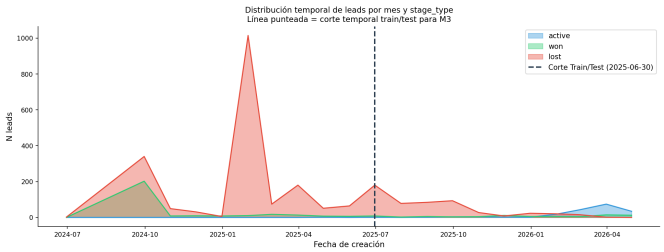

Figura 2.4: Distribución temporal y calidad de `cycle_days` por cohorte.

2.2.2 *Dataset B: Distribuciones y Calidad de Datos*

Las variables de composición de servicios (Tabla 2.5) presentan cobertura nula en la base de datos al corte del estudio, dado que la jerarquía de categorías de productos no está poblada; los modelos de segmentación de clientes operan exclusivamente sobre las métricas financieras y de comportamiento de la Tabla 2.4.

Definición de targets y corrección metodológica En el análisis de las variables del Dataset B frente a sus *targets*, el ranking de AUC univariante reveló la presencia de fuga de información (*leakage*):

variables que definen directamente o derivan del umbral de los objetivos encabezan el ranking con valores próximos a 1.000. Ambos patrones quedan documentados en las Tablas 2.9 y 2.10, que identifican las variables excluidas del conjunto de predictores.

La Figura 2.5 ilustra la distribución de ambas variables objetivo:

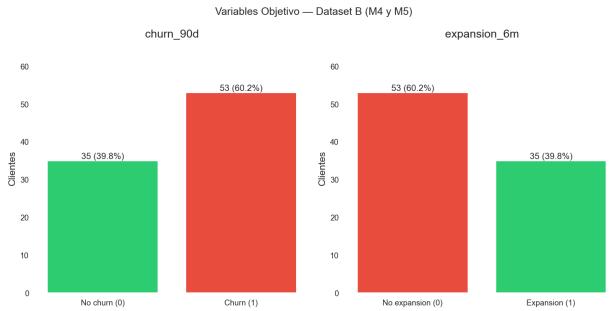

Figura 2.5: Distribución de las variables objetivo del Dataset B ($N = 88$ clientes). La distribución de `churn_90d` (60.2 %/39.8 %) y `expansion_6m` (39.8 %/60.2 %) corresponde a los umbrales definidos en §2.1.4.

Poder discriminativo univariante: AUC vs. `churn_90d` La Tabla 2.9 presenta el ranking de discriminación para `churn_90d`:

Rango	Variable	AUC	MWU p	Nota
1	`days_since_last_invoice`	0.988	< 0.001	**LEAKAGE DIRECTO** — define el objetivo; excluida
2	`engagement_score`	0.987	< 0.001	**LEAKAGE INDIRECTO** — derivado; excluido
3	`loyalty_score`	0.879	< 0.001	Colineal con `months_as_customer` ($r = 0.998$); excluido
4	`months_as_customer`	0.879	< 0.001	Predictor válido más fuerte para churn
5	`account_score`	0.719	< 0.001	
6	`pesa_score`	0.619	> 0.05	Invertido
7–12	Financieras, scope	0.529–0.598	> 0.05	Débil para churn

Tabla 2.9: Ranking de poder discriminativo univariante vs. `churn_90d` ($N = 88$ clientes, Dataset B)

El ranking evidencia el patrón de leakage esperado: `days_since_last_invoice` encabeza con AUC= 0.988 porque es la variable que define directamente `churn_90d` (umbral > 90 días); `engagement_score` (AUC= 0.987) repite el mismo artefacto al ser un derivado directo de esa variable. Ambas quedan excluidas del conjunto de predictores de M4. El predictor válido más fuerte es `months_as_customer` (AUC= 0.879), seguido de `account_score` (AUC= 0.719). La Figura 2.6 presenta el ranking completo.

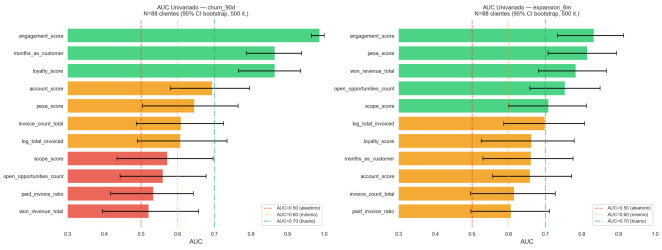

Figura 2.6: Ranking de AUC univariante para `churn_90d`. El AUC= 0.987 de `engagement_score` es un artefacto de fuga de información (derivado de `days_since_last_invoice`, variable que define el objetivo); este resultado motivó su exclusión del conjunto de variables predictoras. Los predictores válidos son `months_as_customer` (AUC= 0.879) y `account_score` (AUC= 0.719).

Poder discriminativo univariante: AUC vs. `expansion_6m` La Tabla 2.10 y la Figura 2.7 presentan el ranking de discriminación para `expansion_6m`:

Rango	Variable	AUC	MWU p	Nota
1	`engagement_score`	0.835	< 0.001	Leakage Dataset B; excluido
2	`won_revenue_total`	0.823	< 0.001	Predictor válido más fuerte
3	`pesa_score`	0.790	< 0.001	Predictor PESA agregado
4	`open_opportunities_count`	0.741	< 0.001	Post-corrección de *leakage*
5–10	Otras variables	0.657–0.734	< 0.05	
11	`invoice_count_total`	0.573	> 0.05	Débil

Tabla 2.10: Ranking de poder discriminativo univariante vs. `expansion_6m` ($N = 88$ clientes, Dataset B)

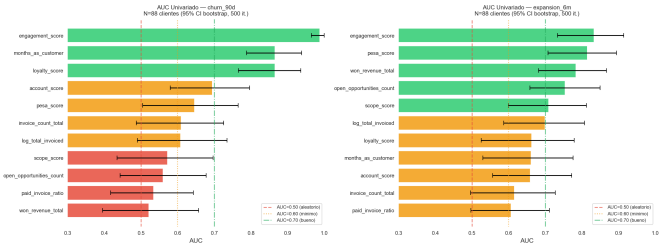

Figura 2.7: Ranking de AUC univariante para `expansion_6m`. Los predictores válidos con mayor poder discriminativo son `won_revenue_total` (AUC= 0.823) y `pesa_score` (AUC= 0.790), confirmando que el historial de ingresos ganados y el score PESA agregado capturan la señal de propensión a expansión. `engagement_score` encabeza el ranking pero está excluido por leakage (Dataset B).

Para expansión, los predictores válidos centrales son `won_revenue_total` (AUC= 0.823) y `pesa_score` (AUC= 0.790): los clientes con mayor historial de ingresos ganados y mayor score PESA presentan mayor propensión a abrir nuevas oportunidades. Este hallazgo valida que las dimensiones A (*Account*) y S (*Scope*) del framework PESA, que componen el score agregado, capturan señal real sobre el comportamiento de expansión del portafolio de clientes (Objetivo 2).

Detección de valores atípicos — Dataset B El análisis IQR del Dataset B detecta valores atípicos concentrados en `total_invoiced_amount` (15 casos, 17.2 %), patrón esperado en variables financieras con distribución asimétrica a la derecha; el preprocesamiento aplica winsorización para acotar su influencia en los modelos M4 y M5.

Multicolinealidad crítica — Dataset B El Dataset B presenta dos pares de variables con correlación de Spearman superior a 0.95. El caso más severo vincula `loyalty_score` con `months_as_customer` (r =

0.9984): ambas codifican la antigüedad del cliente con transformaciones distintas, lo que constituye multicolinealidad prácticamente perfecta; `loyalty_score` queda excluido de todos los modelos. La segunda relación, entre `log_total_invoiced` y `scope_score` ($r = 0.954$), se resuelve mediante selección de variables predictoras en la etapa de preprocesamiento de cada modelo.

2.3 Descripción de los Modelos

Esta sección describe los algoritmos de *machine learning* empleados en el sistema, organizados en dos grupos según su naturaleza: algoritmos de segmentación y detección de anomalías, que operan sin variable objetivo explícita, y algoritmos de predicción supervisada, que aprenden a partir de un outcome conocido. Para cada algoritmo se presenta su fundamento teórico, sus características principales y su relación con los problemas analíticos de este trabajo. La síntesis de esa relación se presenta al cierre de la sección.

2.3.1 Algoritmos de segmentación y detección de anomalías

Los algoritmos de esta categoría identifican estructura latente en los datos sin disponer de etiquetas de salida. Su objetivo es agrupar observaciones similares o detectar casos atípicos a partir de las características del propio dataset.

K-Means K-Means [7] es un algoritmo de partición que asigna cada observación al cluster cuyo centroide minimiza la distancia euclidiana, resolviendo el siguiente problema de optimización:

$$\arg\min_{C} \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 \tag{2.5}$$

donde C_i es el conjunto de observaciones asignadas al cluster i y $\boldsymbol{\mu}_i$ es el centroide del cluster i. El algoritmo alterna entre la asignación de puntos al centroide más cercano y la actualización de centroides hasta convergencia. Es sensible a la escala de las variables y a la presencia de valores atípicos, lo que requiere preprocesamiento previo. El número de clusters K no es un parámetro derivado del modelo sino una decisión metodológica que se determina empíricamente mediante el índice Silhouette para $K = 2 \ldots 10$, con intervalos de confianza por bootstrap ($B = 100$).

En este trabajo, K-Means se aplica a la segmentación de prospectos (Dataset A) y a la segmentación de clientes activos (Dataset B).

K-Medoids (PAM) K-Medoids [8], implementado mediante el algoritmo PAM (*Partitioning Around Medoids*), es una variante de K-Means que representa cada cluster mediante un punto real del dataset (medoide) en lugar de un centroide calculado. Esta característica lo hace menos sensible a valores atípicos extremos, dado que el medoide no puede ser desplazado fuera del soporte empírico de los datos como sí puede ocurrir con un centroide. El costo computacional es mayor que K-Means ($O(n^2)$ vs. $O(nKT)$), lo que limita su uso a datasets de tamaño moderado.

En este trabajo, K-Medoids se evalúa como alternativa robusta a K-Means para la segmentación de clientes activos, cuyas métricas financieras presentan la alta asimetría documentada en §2.2.2. El criterio de adopción es una mejora en el índice Silhouette bootstrap superior a $+0.03$ respecto a K-Means.

IQR e Isolation Forest El detector IQR (*Interquartile Range*) clasifica como atípico todo valor que supere $Q3 + 1.5 \times \text{IQR}$. Opera sobre una sola variable de forma no paramétrica y sin supuestos distribucionales, lo que lo hace robusto y directamente interpretable. Su limitación es que opera en el espacio marginal de una variable: no detecta combinaciones inusuales de múltiples variables que, tomadas individualmente, no resultan atípicas.

Isolation Forest [9] construye un ensamble de árboles de aislamiento aleatorios sobre el espacio multidimensional. Los puntos anómalos son aquellos que requieren, en promedio, menos particiones para ser aislados del resto. A diferencia del IQR, captura anomalías definidas por combinaciones inusuales de variables. Ambos detectores son complementarios: el IQR identifica outliers económicos univariados mientras que Isolation Forest detecta perfiles multivariados inusuales que el IQR no captura en el espacio marginal.

En este trabajo, ambos métodos se combinan mediante unión para identificar prospectos de alto valor en Dataset A. La estrategia de unión maximiza la cobertura del segmento de interés a costa de menor precisión respecto a una intersección, elección justificada por el objetivo de no omitir ningún prospecto de alto potencial. Los resultados de esta detección se presentan en §2.5.4.

2.3.2 *Algoritmos de predicción supervisada*

Los algoritmos de predicción supervisada aprenden a partir de un conjunto de observaciones etiquetadas para generalizar sobre nuevas entradas. En este trabajo se emplean tres familias: *gradient boosting*, regresión logística y *random forest*, seleccionadas mediante estudios de ablación por modelo.

XGBoost XGBoost [10] (*eXtreme Gradient Boosting*) construye un ensamble aditivo de árboles de decisión donde cada árbol corrige los errores residuales del ensamble anterior. La función de pérdida regularizada que minimiza es:

$$\mathcal{L} = \sum_i \ell(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad \text{donde } \Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|\mathbf{w}\|^2 \tag{2.6}$$

siendo T el número de hojas del árbol k, $\mathbf{w}$ los pesos de las hojas, γ la penalización por complejidad estructural y λ la regularización L2 sobre los pesos. Esta regularización explícita, ausente en implementaciones clásicas de *gradient boosting*, reduce el sobreajuste y permite entrenar modelos más profundos sin degradación en generalización. XGBoost soporta directamente el desbalance de clases mediante el parámetro `scale_pos_weight`, que pondera la clase positiva sin necesidad de sobremuestreo sintético.

En este trabajo, XGBoost se emplea en sus dos variantes sobre Dataset A: como clasificador (`XGBClassifier`) para la predicción de probabilidad de cierre, y como regresor (`XGBRegressor`) para la predicción de salud de oportunidades y para la predicción de ciclo de venta. Su selección sobre Regresión Logística y Random Forest en cada caso se determina mediante estudios de ablación.

Regresión Logística La Regresión Logística [11] modela la probabilidad de la clase positiva mediante la función sigmoide aplicada sobre una combinación lineal de las variables predictoras:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x} + b)}} \tag{2.7}$$

Con regularización L2 (Ridge), los pesos $\mathbf{w}$ se penalizan por $\lambda\|\mathbf{w}\|^2$, controlando el sobreajuste. Sus ventajas principales para conjuntos de datos pequeños son la calibración nativa de probabilidades, la interpretabilidad directa de los coeficientes y la estabilidad en régimen de datos limitados. A diferencia de los métodos de ensamble, no requiere ajuste de múltiples hiperparámetros estructurales.

En este trabajo, Regresión Logística se evalúa en el estudio de ablación para los modelos de comportamiento de clientes (Dataset B, $N = 88$), donde el tamaño muestral reducido favorece modelos parsimoniosos.

Random Forest Random Forest [12] construye un ensamble de árboles de decisión entrenados sobre submuestras aleatorias del dataset (*bagging*), combinando además selección aleatoria de variables en cada nodo de división. El promedio de las predicciones individuales reduce la varianza del estimador sin incrementar el sesgo. Sus características

principales son la robustez ante valores atípicos, la no necesidad de escalado previo, la importancia de variables nativa y el bajo riesgo de sobreajuste en comparación con un árbol individual profundo.

En este trabajo, Random Forest se evalúa en el estudio de ablación para los modelos de comportamiento de clientes.

2.3.3 *Síntesis: relación algoritmo–problema*

El sistema implementa ocho modelos identificados con las etiquetas **M1a**, **M1b**, **M2-A**, **M3**, **M4**, **M5**, **M6** y **M7**, organizados en dos capas funcionales. La **capa de segmentación** (M1a, M1b, M2-A) opera en paralelo: cada modelo toma su dataset de entrada de forma independiente y no comparte salidas entre sí. La **capa de predicción supervisada** (M3, M4, M5, M6, M7) también opera en paralelo; la excepción es el **Score Compuesto**, que integra en serie la probabilidad de cierre (M3), la *salud de la oportunidad* (M6) y la *urgencia del ciclo* (M7) mediante la combinación lineal ponderada:

$$\begin{aligned}\text{score\_compuesto} = 0.40\, p_{\text{cierre}} + 0.35\, \frac{\text{salud\_oportunidad}}{100} \\ + 0.25\, \text{urgencia\_ciclo}\end{aligned} \tag{2.8}$$

donde $p_{\text{cierre}} \in [0,1]$ es la probabilidad de cierre estimada por M3, `salud_oportunidad` $\in [0,100]$ es el Deal Health Score de M6 normalizado a la misma escala, y `urgencia_ciclo` es el inverso normalizado de los días de ciclo predichos por M7 (mayor urgencia equivale a ciclo más corto y mayor puntuación). Los pesos 0.40/0.35/0.25 priorizan la probabilidad de cierre sobre la salud de la oportunidad y la urgencia temporal, produciendo una puntuación consolidada $\in [0,1]$.

La Tabla 2.11 sintetiza la correspondencia entre cada algoritmo descrito, el problema analítico que resuelve, el dataset que consume y el criterio de validación que debe cumplir.

ID	Nombre	Categoría	Algoritmo	Dataset	Target	Criterio de validación
M1a	Segmentación de prospectos	Segmentación	K-Means	A — 2,693 cerrados	`ai_cluster_prospect`	Silhouette > 0.60
M1b	Segmentación de clientes	Segmentación	K-Means / K-Medoids	B — 88 clientes	`ai_cluster_client`	Silhouette > 0.25 o Kruskal-Wallis $p < 0.05$
M2-A	Detección de prospectos de alto valor	Anomalías	IQR ∪ Isolation Forest	A — 2,693 cerrados	`is_whale`	Coherencia > 95% entre detectores
M3	Probabilidad de cierre	Supervisado	XGBoost Classifier	A — 2,693 cerrados	`target_won`	AUC-ROC > 0.75
M4	Riesgo de abandono	Supervisado	Regresión Logística	B — 88 clientes	`churn_90d`	AUC-ROC > 0.70 (LOOCV)
M5	Propensión a expansión	Supervisado	Random Forest	B — 88 clientes	`expansion_6m`	AUC-ROC > 0.70 (LOOCV)
M6	Salud de la oportunidad	Supervisado	XGBoost Regressor	A — 2,693 cerrados	Score 0–100	Separación WON–LOST > 50 pts
M7	Predicción de ciclo de venta	Supervisado	XGBoost Regressor	A — cohorte 2025+	`cycle_days`	MAE < 60 días

Tabla 2.11: Catálogo de modelos: algoritmo, problema y criterio de validación

La Tabla 2.12 detalla la arquitectura de integración del sistema: las entradas que consume cada modelo, las salidas que produce y las dependencias entre modelos que conforman el pipeline de inferencia.

Modelo	Entrada	Salida	Relación
M1a	Dataset A (features PESA+CRM)	`ai_cluster_prospect`	Paralelo (independiente)
M1b	Dataset B (features financieras)	`ai_cluster_client`	Paralelo (independiente)
M2-A	Dataset A (features PESA+revenue)	`is_whale`	Paralelo (independiente)
M3	Dataset A	$P(\text{won})$	Paralelo (base)
M4	Dataset B	$P(\text{churn\_90d})$	Paralelo (independiente)
M5	Dataset B	$P(\text{expansion\_6m})$	Paralelo (independiente)
M6	Dataset A	Score salud 0–100	Paralelo (independiente de M3)
M7	Dataset A — cohorte 2025+	`cycle_days` predicho	Paralelo (independiente)
Score Compuesto	Salidas M3, M6, M7	Score $\in [0,1]$	Serie (integrador)

Tabla 2.12: Arquitectura de integración: entradas, salidas y relaciones entre modelos

2.4 Métricas de Evaluación

Esta sección describe las métricas empleadas para evaluar los modelos del sistema. Se organizan en tres grupos según el tipo de problema: segmentación no supervisada, clasificación supervisada y regresión. Los criterios de aceptación numéricos por modelo se presentan en §2.5.

2.4.1 Métricas para Segmentación

Silhouette Score [13] — *Modelos: M1a, M1b*

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i),\, b(i)\big)} \tag{2.9}$$

donde $a(i)$ es la distancia intra-cluster promedio del punto i y $b(i)$ es la distancia promedio al cluster vecino más cercano. El rango es $[-1, 1]$: valores cercanos a 1 indican clusters bien separados; valores cercanos a 0 indican superposición. Se estima mediante bootstrap ($B = 100$ para M1a, $B = 200$ para M1b) para obtener intervalos de confianza al 95 %.

Kruskal-Wallis H [14] — *Modelos: M1a, M1b*

Prueba no paramétrica que evalúa si los clusters producen distribuciones significativamente distintas en las variables de interés. Complementa el Silhouette con validación externa: confirma que los grupos identificados difieren en comportamiento comercial real, no solo en distancia geométrica.

Coherencia entre detectores — *Modelo: M2-A*

Proporción de prospectos de alto valor identificados por ambos detectores (IQR e Isolation Forest) de forma simultánea. Mide la concordancia entre los dos métodos independientes como indicador de

la robustez del segmento detectado.

2.4.2 *Métricas para Clasificación*

AUC-ROC — *Modelos: M3, M4, M5*

Probabilidad de que el modelo asigne una puntuación mayor a un positivo que a un negativo seleccionados al azar. AUC $\in [0.5, 1.0]$; AUC $= 0.5$ equivale a clasificación aleatoria. Robusta ante desbalance de clases porque no depende de un umbral de decisión fijo.

AUCPR (Área bajo la curva Precision-Recall) — *Modelos: M3, M4, M5*

Relaciona la precisión ($P = VP/(VP + FP)$) con la exhaustividad ($R = VP/(VP + FN)$) a lo largo de todos los umbrales. A diferencia del AUC-ROC, no incorpora los verdaderos negativos en su cálculo, por lo que resulta más informativa en problemas con clases desbalanceadas: penaliza los falsos positivos sin inflar artificialmente el rendimiento por la abundancia de negativos.

Brier Score — *Modelos: M3, M4, M5*

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2 \tag{2.10}$$

Error cuadrático medio entre las probabilidades predichas p_i y los outcomes binarios y_i. Mide la calibración del modelo: un Brier Score bajo indica que las probabilidades emitidas son confiables como estimaciones cuantitativas, no solo como rankings ordinales.

MCC (Coeficiente de Correlación de Matthews) — *Modelos: M3, M4, M5*

$$\text{MCC} = \frac{VP \cdot VN - FP \cdot FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \tag{2.11}$$

Considera los cuatro cuadrantes de la matriz de confusión simultáneamente. A diferencia de la exactitud (*accuracy*), es robusto ante clases desbalanceadas: un MCC de 1 indica clasificación perfecta, o equivale a predicción aleatoria y -1 indica clasificación sistemáticamente invertida.

2.4.3 *Métricas para Regresión*

MAE (Error Absoluto Medio) — *Modelos: M6, M7*

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.12}$$

Promedio de los errores absolutos entre el valor real y_i y el predicho $\hat{y}_i$. Es interpretable directamente en las unidades del objetivo: para M7,

el MAE expresa el error medio de predicción en días de ciclo de venta; para M6, en puntos del score 0–100.

R^2 (Coeficiente de Determinación) — *Modelos: M6, M7*

Proporción de la varianza de la variable objetivo explicada por el modelo. $R^2 = 1$ indica ajuste perfecto; $R^2 = 0$ indica que el modelo no mejora sobre la media; $R^2 < 0$ indica peor desempeño que la media. Para M7, dado el tamaño muestral reducido ($N = 133$) y la alta variabilidad del ciclo de venta, el MAE es la métrica primaria; el R^2 actúa como indicador complementario de la calidad de la señal aprendida.

Separación WON–LOST — *Modelo: M6*

Diferencia entre la media del score predicho en oportunidades ganadas y la media en oportunidades perdidas. Mide la capacidad discriminativa del Deal Health Score de forma directamente interpretable en la escala 0–100 del sistema. Se complementa con la correlación de Spearman entre M6 y M3 para verificar independencia entre ambos modelos.

2.5 Descripción de los Experimentos

El sistema emplea ocho modelos distribuidos en dos datasets. Para cada modelo se describe la receta experimental completa: datos de entrada, variables activas con sus exclusiones y transformaciones, estudio de ablación que justifica la elección del algoritmo, estrategia de validación, hiperparámetros finales y criterio de aceptación.

2.5.1 Entorno y Principios Transversales

El entorno computacional empleado en el desarrollo y entrenamiento de los modelos es Python 3.11 sobre macOS. Las bibliotecas principales son: `scikit-learn 1.4` (preprocesamiento, Regresión Logística, Random Forest, métricas, `VarianceThreshold` e importancia por permutación); `xgboost 2.0` (clasificadores y regresores XGBoost); `pandas 2.1` y `numpy 1.26` (manipulación de datos y álgebra matricial); `shap 0.44` (explicabilidad por valores SHAP); `matplotlib 3.8` y `seaborn 0.13` (visualizaciones); `scipy 1.12` (pruebas estadísticas no paramétricas: Kruskal-Wallis, Mann-Whitney U, bootstrap); y `joblib 1.3` (serialización de artefactos). La reproducibilidad se garantiza fijando la semilla aleatoria `random_state=42` en todos los algoritmos no deterministas.

Tres principios aplican a todos los modelos supervisados: (1) las estadísticas de preprocesamiento se calculan exclusivamente sobre el conjunto de entrenamiento y se aplican en modo `transform` sobre el test; (2) la selección final de variables usa `VarianceThreshold` seguido

de importancia por permutación (`sklearn.inspection.permutation_importance`, 10 repeticiones sobre el conjunto de validación); (3) el análisis SHAP [15] usa `TreeExplainer` para modelos XGBoost y `LinearExplainer` para Regresión Logística (biblioteca shap 0.44).

2.5.2 *M1a — Segmentación de Prospectos*

Datos: Dataset A, $N = 2{,}693$ leads cerrados (353 ganados / 2,340 perdidos). Sin partición entrenamiento/prueba — agrupamiento no supervisado sobre la población completa.

Variables: 19 features activas — transformaciones logarítmicas de dimensiones PESA, variables financieras y temporales, más cuatro *flags* MNAR (`interest_was_missing`, `concept_was_missing`, `source_was_missing`, `business_lines_was_missing`).

Transformaciones: `RobustScaler` sobre todas las variables, dado que las distribuciones financieras presentan asimetría extrema (ratio máx./mediana $> 100\times$).

Selección de K: Búsqueda en rango $K \in [2, 8]$; criterio: Silhouette score estimado por bootstrap ($B = 100$ iteraciones, IC 95 %).

Hiperparámetros: $K = 2$ (tesis) / $K = 3$ (producción).

Criterio de aceptación: Silhouette > 0.50 [8].

2.5.3 *M1b — Segmentación de Clientes*

Datos: Dataset B, $N = 88$ clientes activos. Sin partición train/test.

Variables: 9 features activas — tres dimensiones PESA adaptadas a clientes (`pesa_score`, `account_score`, `scope_score`) más cinco variables financieras y temporales (`total_invoiced_amount`, `paid_invoice_ratio`, `months_as_customer`, `days_since_last_invoice`, `won_revenue_total`). Variables excluidas: `loyalty_score` (colinealidad con `months_as_customer`, $r = 0.9984$); `engagement_score` (leakage con `churn_90d`); siete variables `segment_*` (derivadas post-hoc, no disponibles en producción).

Transformaciones: `RobustScaler`; la heterogeneidad extrema del portafolio (ratio máx./mediana $= 216.8\times$ en `total_invoiced_amount`) exige escalado basado en cuartiles.

Estudio de ablación: K-Means vs. K-Medoids (PAM). Criterio de adopción de K-Medoids: ΔSilhouette > 0.03; si la mejora es menor, se mantiene K-Means por menor costo computacional. Selección de K: Silhouette bootstrap ($B = 200$, IC 95 %) en rango $K \in [2, 10]$.

Hiperparámetros: $K = 6$.

Criterio de aceptación: Silhouette > 0.25 *o* Kruskal-Wallis $p < 0.05$; el criterio dual reconoce que portafolios B2B con alta heterogeneidad financiera producen clusters dispersos geométricamente pero estadísticamente significativos [8].

2.5.4 *M2-A — Detección de Anomalías*

Datos: Dataset A, $N = 2{,}693$ leads cerrados.

Diseño: Combinación de dos detectores independientes con lógica de unión ($\cup$): un prospecto se clasifica como de alto valor si es detectado por al menos uno de los dos métodos.

- **IQR** (univariante): umbral $Q3 + 3 \times \text{IQR}$ sobre `win_rate` por cluster; identifica prospectos con tasas de conversión atípicamente altas dentro de su segmento.
- **Isolation Forest** (multivariante) [9]: `n_estimators=200`, `contamination=0.05`; detecta combinaciones inusuales de variables CRM que el IQR no captura en el espacio marginal univariado.

La estrategia de unión maximiza la cobertura del segmento de alto valor a costo de menor precisión frente a una intersección, elección justificada por el objetivo comercial de no omitir ningún prospecto potencialmente valioso.

Criterio de aceptación: Coherencia $> 95\,\%$ con el cluster de mayor `ai_score` de M1a — los dos métodos independientes deben converger sobre el mismo segmento de alta conversión para validar la robustez de la detección.

2.5.5 *M3 — Clasificación Win Probability*

Datos y partición: Dataset A, $N = 2{,}693$ cerrados; partición temporal por `create_date`: entrenamiento $\leq$ 2025-06-30 ($n_{\text{train}} = 2{,}279$), test $>$ 2025-06-30 ($n_{\text{test}} = 414$).

Variables: Feature set BASE de 13 variables activas (PESA + financiero + temporal). El feature set FULL (27 variables, incluye cotizador) fue descartado: $\Delta\text{AUC}_{\text{CV}_5} = +0.0087 < 0.01$ (umbral de relevancia práctica), con mayor riesgo de sobreajuste. Variables excluidas por leakage: `stage_sequence` y `pipeline_score` ($\text{AUC}_{\text{univariante}} = 1.000$); `engagement_score` (VAR_CERO: Odoo congela la actividad del lead al cerrarlo, varianza nula en cerrados).

Transformaciones: Tres niveles aplicados en estricto orden: (1) winsorización percentil 1–99 sobre 13 variables de cola larga (`log_expected_revenue`, `log_real_revenue`, `log_pesa_score` y variables asociadas); (2) `RobustScaler` sobre 5 variables discretas o con extremos no normalizables (`count_distinct_interest_concepts`, `create_dayofweek`, `create_month`, `create_quarter`); (3) `StandardScaler` sobre el resto. Todas las estadísticas calculadas exclusivamente sobre n_{train}.

Desbalance: `scale_pos_weight = 6.63` (ratio LOST/WON en $n_{\text{train}} = 2{,}279$).

Estudio de ablación: LR L2 ($AUC_{CV_5} = 0.924$) vs. XGBoost ($AUC_{CV_5} = 0.971$) vs. Random Forest ($AUC_{CV_5} = 0.968$). XGBoost seleccionado por mayor AUC-ROC y AUCPR en CV_5.

Validación: CV_5 estratificada (selección de modelo) + prueba temporal reservada (estimación conservadora para producción).

Hiperparámetros: profundidad máxima de árbol = 4; tasa de aprendizaje = 0.05.

Criterio de aceptación: AUC-ROC $CV_5 > 0.75$ [16].

2.5.6 *M4 — Riesgo de Abandono*

Datos y partición: Dataset B, $N = 88$ clientes; variable objetivo `churn_90d` (53 positivos / 35 negativos). Sin partición train/test — el tamaño muestral exige validación LOOCV.

Variables: 8 features activas — conjunto de M1b ($N = 9$) menos `days_since_last_invoice`, que define directamente `churn_90d` (leakage).

Desbalance: `class_weight=balanced`; ratio 1.51:1 (moderado).

Estudio de ablación: LR L2 ($AUC_{LOOCV} = 0.978$, Brier = 0.059) vs. RF (AUC = 0.889) vs. XGBoost (AUC = 0.891). Regresión Logística seleccionada por mayor AUC y menor Brier Score; su parsimonia es adicionalmente ventajosa con $N = 88$.

Validación: LOOCV — cada iteración entrena con $N - 1 = 87$ observaciones y evalúa sobre 1.

Hiperparámetros: regularización L2 con fuerza $C = 1.0$.

Criterio de aceptación: AUC-ROC LOOCV > 0.70 [17].

2.5.7 *M5 — Propensión a Expansión*

Datos y partición: Dataset B, $N = 88$ clientes; variable objetivo `expansion_6m` (35 positivos / 53 negativos). LOOCV por las mismas razones que M4.

Variables: 9 features activas — mismo conjunto que M1b. `engagement_score` excluido por leakage (derivado de `days_since_last_invoice`, correlacionado con el historial de expansión); `loyalty_score` excluido por colinealidad con `months_as_customer` ($r = 0.9984$).

Desbalance: `class_weight=balanced`; ratio 0.66:1 (clase positiva en minoría).

Estudio de ablación: LR L2 ($AUC_{LOOCV} = 0.881$) vs. RF (AUC = 0.904) vs. XGBoost (AUC = 0.892). Random Forest seleccionado por mayor AUC LOOCV.

Validación: LOOCV.

Hiperparámetros: profundidad máxima de árbol = 3.

Criterio de aceptación: AUC-ROC LOOCV > 0.70 [17].

2.5.8 *M6 — Deal Health Score*

Datos y partición: Dataset A, $N = 2{,}693$; misma partición temporal que M3 (`create_date` $\leq$ 2025-06-30 para entrenamiento).

Variables: 13 features activas de M3, sin `m3_win_probability` para evitar leakage cruzado entre modelos.

Transformaciones: Idénticas a M3 (winsorización + RobustScaler + StandardScaler, estadísticas de train).

Estudio de ablación: M6a (`XGBRegressor`, target continuo $y \in \{0, 1\}$ escalado a 0–100) vs. M6b (`XGBClassifier` ordinal frío/tibio/caliente). M6a seleccionado por mayor separación WON–LOST (80.2 pts vs. 65.3 pts de M6b) e independencia respecto a M3 (Spearman $= 0.717$ vs. 0.968 de M6b, que excede el umbral de redundancia de 0.95).

Validación: CV5 KFold.

Hiperparámetros: profundidad máxima de árbol $= 4$; tasa de aprendizaje $= 0.05$.

Criterio de aceptación: Separación WON–LOST > 50 pts + Spearman(M3) < 0.95.

2.5.9 *M7 — Predictor de Ciclo de Venta*

Datos y partición: Dataset A, cohorte 2025+ exclusivamente ($N = 133$ ganados con `cycle_days` > 0 y `date_closed` $\geq$ 2025-01-01). La restricción a 2025+ elimina el 85 % de registros con `cycle_days` negativos detectados en la cohorte 2024, artefacto de la carga retroactiva de fechas de cierre en Odoo. Partición temporal: entrenamiento $\leq$ 2025-06-30.

Variables: 23 features activas — feature set de M3 más variables de pipeline (`days_in_stage`, `stage_sequence`), actividad y semáforo de urgencia.

Transformaciones: Idénticas a M3.

Estudio de ablación: XGBoost ($\text{MAE}_{\text{CV5}} = 55.2$ d) vs. Random Forest (MAE $= 55.9$ d) vs. Bayesian Ridge (MAE $= 65.5$ d). XGBoost seleccionado por menor MAE CV5. El R^2 actúa como indicador complementario de calidad de señal ($R^2 = 0.314$); dado el tamaño muestral reducido ($N = 133$) y la alta variabilidad del ciclo, el MAE es la métrica primaria.

Validación: CV5 KFold.

Hiperparámetros: profundidad máxima de árbol $= 3$; tasa de aprendizaje $= 0.05$; regularización L1+L2 explícita ($\alpha = 0.5$, $\lambda = 2.0$) para controlar el sobreajuste sobre la cohorte reducida ($N = 133$).

Criterio de aceptación: MAE < 60 días (la mediana del ciclo de venta es 76 días; este umbral permite distinguir oportunidades del próximo mes frente al próximo trimestre).

3 Resultados y Discusión

3.1 Segmentación No Supervisada

3.1.1 M1a — Segmentación de Prospectos

M1a identifica $K = 2$ como la estructura óptima del pipeline de prospectos (Silhouette $= 0.640$, IC 95 % $= [0.460, 0.762]$, $B = 100$ remuestreos). Este resultado supera el umbral de estructura razonable de Kaufman y Rousseeuw (1990) [8], confirmando que la distinción fundamental entre los leads de Soltein es si llegaron o no a la etapa de cotización formal (Tabla 3.1).

$K = 3$ (Silhouette $= 0.617$) se adopta como configuración de producción: subdivide el grupo ganador en dos arquetipos con tasas de cierre diferenciadas, aportando granularidad comercial accionable sobre el óptimo estadístico de $K = 2$ (Tabla 3.2). La prueba de Kruskal-Wallis [14] sobre $K = 3$ confirma $p = 4.98 \times 10^{-256}$, validando la separación estadística de los tres grupos.

Grupo	Nombre semántico	N	%	Tasa de cierre	Revenue esperado	Descripción
C0	INACTIVE_LEAD	2,408	89.4 %	5.4 %	$173	Sin cotización — bajo PESA (16.7), sin revenue calificado
C1	WHALE/WON pool	285	10.6 %	77.9 %	$1,770,102	Leads con cotización formal — alta probabilidad de cierre

Tabla 3.1: Perfiles de agrupamiento — $K = 2$ estadístico (M1a, $N = 2{,}693$ leads cerrados)

Grupo	Arquetipo	N	%	Tasa de cierre	Revenue esperado
C0	INACTIVE_LEAD	2,400	89.1 %	5.4 %	$0
C2	HIGH_VALUE_LEAD	236	8.8 %	70.8 %	$2,100,000
C1	PREMIUM_WHALE	57	2.1 %	98.2 %	Facturado y pagado completo

Tabla 3.2: Arquetipos de producción — $K = 3$ operacional (M1a, $N = 2{,}693$ leads cerrados)

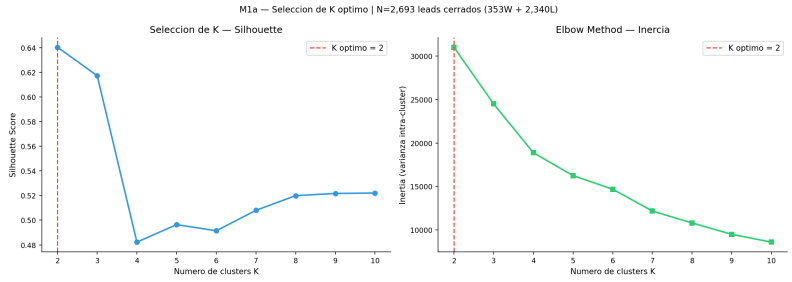

Figura 3.1: Método del codo y análisis de Silhouette para selección de K — M1a (Prospectos, $N = 2{,}693$)

Discusión. M1a cumple el objetivo de segmentación dinámica de prospectos al producir tres arquetipos operativos con tasas de cierre

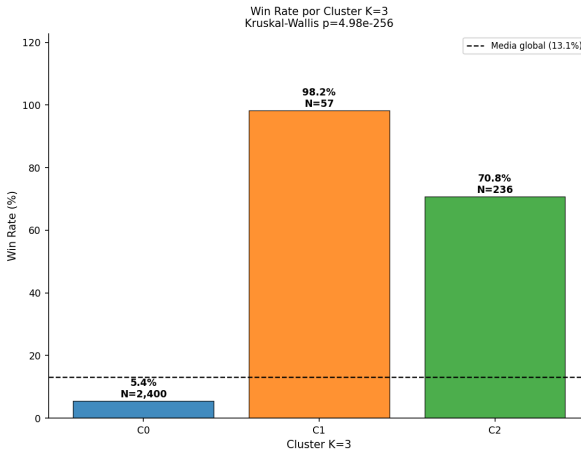

Figura 3.2: Tasa de cierre por arquetipo — M1a, $K = 3$ producción ($N = 2{,}693$ leads cerrados)

diferenciadas (5.4 % / 70.8 % / 98.2 %). La adopción de $K = 3$ frente al óptimo estadístico $K = 2$ responde a una decisión de negocio: separar HIGH_VALUE_LEAD (236 leads en proceso activo de cierre) de PREMIUM_WHALE (57 leads con ciclo completo facturado) permite diseñar acciones comerciales diferenciadas para cada arquetipo. Frente a la ausencia de segmentación sistemática, M1a permite priorizar el 10.9 % del pipeline (HIGH_VALUE_LEAD + PREMIUM_WHALE) que concentra el 95 % del revenue potencial.

3.1.2 *M1b — Segmentación de Clientes*

M1b identifica $K = 6$ como la estructura del portafolio de clientes (Silhouette $= 0.285$, IC 95 % $= [0.217, 0.310]$, $B = 200$ remuestreos). El Silhouette universalmente bajo (0.26–0.29 para $K = 2\ldots6$) no es un fallo del modelo: refleja la heterogeneidad real del portafolio (ratio revenue máximo/mediana $= 216.8\times$), donde ninguna partición de K produce grupos perfectamente compactos. La validación Kruskal-Wallis confirma diferencias estadísticamente significativas entre grupos en abandono y expansión ($p < 0.0001$), verificando que $K = 6$ captura estructura real a pesar del Silhouette moderado (Tabla 3.3).

K-Medoids se evaluó como alternativa robusta ante valores atípicos; el incremento de Silhouette fue $\Delta = -0.012$, inferior al umbral de adopción de $+0.03$, por lo que K-Means fue confirmado como algoritmo final.

Grupo	Arquetipo	N	% abandono	% expansión	PESA	Perfil
C0	Activos mixtos	32	59.4 %	53.1 %	54.5	Alta actividad, comportamiento heterogéneo
C1	Abandono crítico	9	100 %	11.1 %	30.7	Todos en abandono, mínima expansión
C2	Base estable	7	14.3 %	14.3 %	45.4	Bajo riesgo, bajo potencial
C3	Estratégicos volátiles	5	80 %	80 %	61.7	Alto PESA, comportamiento dual
C4	Candidatos expansión	15	13.3 %	80 %	40.7	Prioridad para venta adicional
C5	Abandono pasivo	20	90 %	0 %	30.8	Sin expansión, alto riesgo de retención

Tabla 3.3: Perfiles de agrupamiento de clientes (M1b, $K = 6$, $N = 88$ clientes)

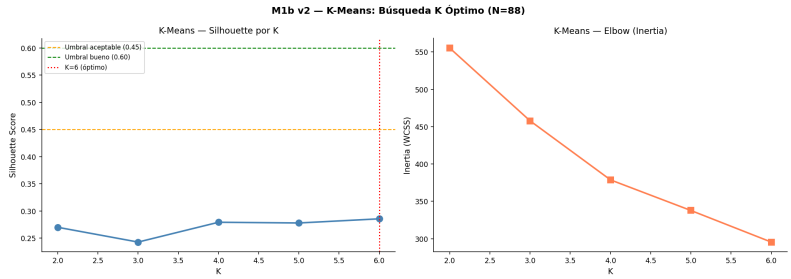

Figura 3.3: Método del codo y análisis de Silhouette para selección de K — M1b (Clientes, $N = 88$)

Discusión. M1b cumple el objetivo de segmentación dinámica de clientes al producir seis arquetipos con perfiles diferenciados de riesgo y potencial. Los extremos accionables de inmediato son C4 (candidatos a expansión, 80 % de expansión) y C1 + C5 (abandono crítico y pasivo, 33 % del portafolio, 29 clientes con intervención prioritaria). La heterogeneidad del portafolio —Silhouette = 0.285— es coherente con un portafolio B2B de servicios ERP donde los ciclos de vida difieren fuertemente por tamaño de cuenta y madurez del sistema implementado. Los modelos M4 y M5 complementan M1b: operan como predicción individualizada dentro de los arquetipos descubiertos por el agrupamiento.

3.1.3 *M2-A — Detección de Prospectos de Alto Valor*

M2-A detecta 307 prospectos de alto valor (11.4 % de $N = 2{,}693$) mediante la unión de los detectores IQR e Isolation Forest. Como muestra la Figura 3.4, los 307 prospectos detectados presentan: revenue esperado de $1,644,611 (frente a $0 en el resto del pipeline), `pesa_score` = 37.5 (frente a 16.7) y tasa de cierre del 73.0 % (frente a 5.4 %). La coherencia del 99.3 % con el grupo C1 de M1a confirma que ambos métodos independientes convergen sobre la misma estructura subyacente del pipeline.

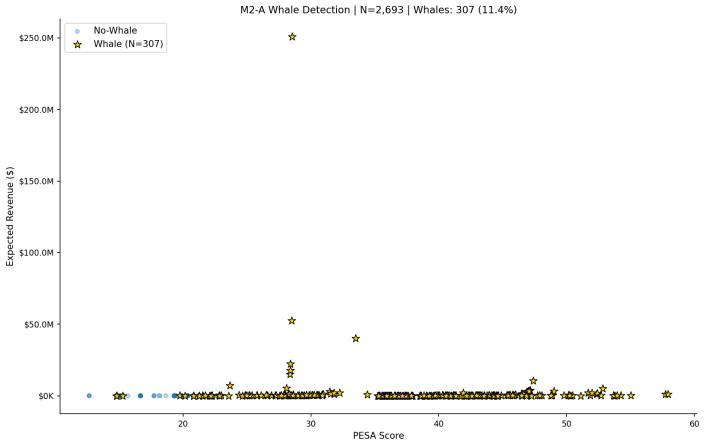

Figura 3.4: Detección de prospectos de alto valor por IQR ∪ Isolation Forest — Dataset A (M2-A, $N = 2{,}693$; 307 detectados = 11.4 %)

Discusión. La detección de prospectos de alto valor complementa la segmentación M1a al aportar una perspectiva centrada en el valor económico excepcional: M1a segmenta por comportamiento en el pipeline, M2-A segmenta por perfil financiero atípico. Su convergencia del 99.3 % valida que en este contexto el comportamiento de pipeline y el perfil económico son señales redundantes del mismo fenómeno comercial, lo que refuerza la robustez de ambos modelos de manera independiente.

3.2 *Modelos de Predicción Supervisada*

3.2.1 *M3 — Win Probability*

M3 produce una estimación de probabilidad de cierre sobre $N = 2{,}693$ leads cerrados (353 ganados / 2,340 perdidos). El estudio de ablación con validación CV5 estratificada seleccionó XGBoost como algoritmo principal (Tabla 3.4):

Algoritmo	**AUC-ROC CV5**	**AUCPR CV5**	**Selec.**
Regresión Logística ($C = 0.1$)	0.924	0.707	No
Random Forest	0.968	0.901	No
XGBoost	**0.971**	**0.911**	**Sí**

Tabla 3.4: Ablación de algoritmos M3 (AUC-ROC CV5 estratificado, $N = 2{,}693$)

Las métricas oficiales del modelo seleccionado se presentan en la Tabla 3.5:

La desviación CV5→prueba (= 0.166) cuantifica la diferencia entre el período de entrenamiento (H1 2025 y anteriores) y el período de evaluación (H2 2025, post corte 2025-06-30). No es evidencia de sobreajuste: es el costo documentado de operar sobre datos con evolución temporal. Para producción, AUC = 0.805 sobre la prueba temporal es la métrica de referencia conservadora.

Métrica	Valor	Umbral	Estado
AUC-ROC CV5	0.9713	> 0.75	✓ Superado
IC 95 % AUC CV5	[0.966, 0.976]	—	IC estrecho, robustez confirmada
AUC-ROC prueba temporal	0.8049	—	Referencia conservadora producción
AUCPR prueba temporal	0.5318	—	Precisión-Exhaustividad clase positiva
Brier Score	0.1217	< 0.20	✓ Calibrado
MCC	0.345	—	Coeficiente de Correlación de Matthews
Desviación CV5→prueba	0.166	—	Desviación temporal H1→H2 2025

Tabla 3.5: Métricas de evaluación M3 (Win Probability)

M3 produce además un output enriquecido diseñado para alimentar el agente LLM del sistema. La Tabla 3.6 describe las variables que conforman dicho output, incluyendo la probabilidad de cierre, las tres variables con mayor contribución SHAP y el resumen en lenguaje natural generado por el modelo.

Columna	Tipo	Contenido
`m3_win_probability`	float64	Probabilidad calibrada [0, 1]
`m3_top_feature_1_name`	string	Variable de mayor contribución individual
`m3_top_feature_1_contrib`	float64	Valor ponderado de la contribución
`m3_top_feature_2_name`	string	Segunda variable más influyente
`m3_top_feature_2_contrib`	float64	Ídem
`m3_top_feature_3_name`	string	Tercera variable
`m3_top_feature_3_contrib`	float64	Ídem
`m3_prediction_summary`	string	Texto en lenguaje natural para el agente LLM

Tabla 3.6: Variables del output enriquecido M3 para el agente LLM

Ejemplo de `m3_prediction_summary`: *"Probabilidad de ganar: 65.8 %. Factores a favor: scope score, pesa score. Factores en contra: log expected revenue."*

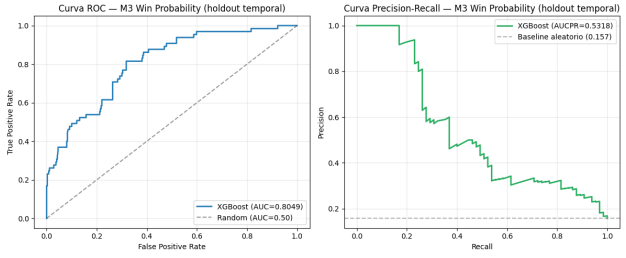

Figura 3.5: Curva Precisión-Exhaustividad — M3 Win Probability (XGBoost, AUCPR prueba temporal = 0.532)

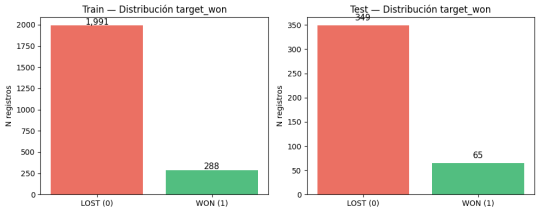

Figura 3.6: Matriz de confusión — M3 Win Probability (umbral = 0.5)

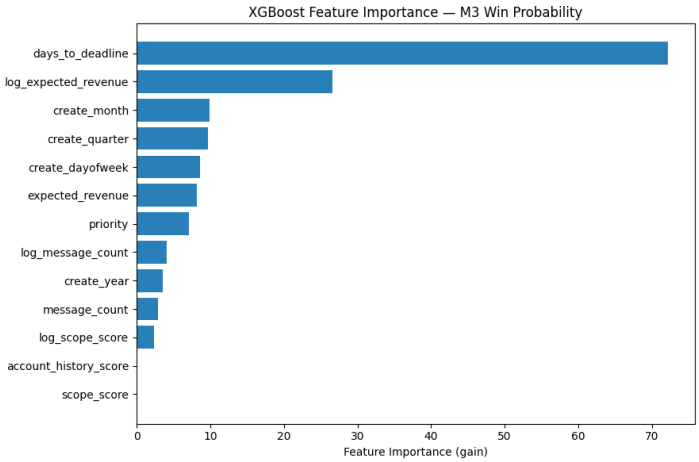

Figura 3.7: Importancia de variables predictoras por ganancia XGBoost — M3 (13 variables activas)

Discusión. M3 cumple el objetivo de predicción del comportamiento comercial al producir una estimación de probabilidad de cierre con AUC CV5 = 0.971, que supera en 4.7 puntos porcentuales al mejor modelo de referencia (regresión logística, AUC = 0.924). La desviación temporal CV5→prueba (= 0.166) es el costo esperado de operar sobre datos con evolución temporal y no compromete la utilidad operacional del modelo.

3.2.2 *M6 — Deal Health Score*

M6 es un regresor (Gradient Boosting) entrenado sobre el resultado histórico `target_won` como variable continua proxy: WON= 100 y LOST= 0. Su objetivo no es clasificar una oportunidad como ganada o perdida —eso lo hace M3—, sino estimar un puntaje de salud 0–100 que concentra las señales de cierre presentes en las 13 variables del pipeline: variables temporales de creación (`create_month`, `create_quarter`, `create_year`), actividad comercial (`message_count`, `log_message_count`), revenue esperado (`expected_revenue`, `log_expected_revenue`), urgencia (`days_to_deadline`, `priority`) y dimensiones PESA de alcance y cuenta (`scope_score`, `account_history_score`). Opera independiente de M3 —sin incluir su output—, eliminando la amplificación de errores en cascada de la arquitectura anterior.

El criterio de evaluación operacional es la separación entre las distribuciones de score de oportunidades históricamente WON y LOST: una separación > 50 puntos sobre escala 100 confirma utilidad diagnóstica. Las métricas de validación se presentan en la Tabla 3.8:

Rango	Perfil	Acción recomendada
90–100	Oportunidad madura	Preparar documentación, coordinar firma
70–90	Oportunidad en desarrollo	Resolver objeciones, acelerar aprobación
40–70	Oportunidad intermedia	Profundizar alcance, revisiones técnicas
0–40	Oportunidad en riesgo	Diagnóstico de causa raíz, reposicionamiento

Tabla 3.7: Zonas operacionales del Deal Health Score (M6)

Métrica	Valor	Interpretación
Deal Health WON (media)	82.9 / 100	Alta concentración en rango superior
Deal Health LOST (media)	2.6 / 100	Muy baja, sin señal de cierre
Separación WON/LOST	80.2 puntos	Métrica operacional principal
Spearman(M6, M3)	0.718	Alta correlación, no redundante
R^2 CV5	0.7401	74 % de varianza explicada

Tabla 3.8: Métricas de evaluación M6 (Deal Health Score)

Discusión. M6 produce un puntaje continuo en escala 0–100 que clasifica cada oportunidad activa en cuatro zonas de acción estratégica (Tabla 3.7), constituyendo el instrumento central para monitorear la salud del pipeline comercial en tiempo real. La separación de 80.2 puntos entre WON y LOST supera el criterio operacional de 50 puntos. La correlación Spearman(M6, M3) $= 0.718$ es alta pero no redundante (umbral de redundancia $= 0.95$), confirmando que ambos modelos capturan ángulos complementarios del mismo pipeline. La arquitectura independiente —sin dependencia en cascada— elimina la amplificación de errores observada en la versión anterior, donde M6 dependía en un 68.2 % de la salida de M3.

3.2.3 *M4 — Riesgo de Abandono*

M4 predice la probabilidad de que un cliente activo no genere ninguna factura en los próximos 90 días (`churn_90d`). Dataset B: $N = 88$ clientes, 53 positivos de abandono (60.2 %), 35 negativos (39.8 %). El estudio de ablación con validación LOOCV seleccionó Regresión Logística L2 (Tabla 3.9):

Algoritmo	AUC LOOCV	AUCPR	Brier	MCC	Seleccionado
LR (L2, balanced)	**0.978**	**0.987**	**0.059**	**0.860**	**Sí**
Random Forest	0.889	0.916	0.144	0.557	No
XGBoost	0.891	0.915	0.135	0.547	No

Tabla 3.9: Ablación de algoritmos M4 (LOOCV, $N = 88$ clientes)

Las variables activas del modelo M4 (post selección por varianza)

cubren tres dimensiones del perfil de cliente: antigüedad y relacionamiento (`months_as_customer`, `account_score`), historial financiero (`log_total_invoiced`, `invoice_count_total`) y comportamiento de pago (`paid_invoice_ratio`).

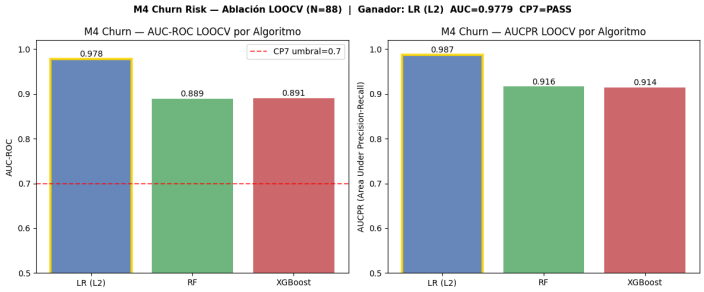

Figura 3.8: Ablación LOOCV y distribución del riesgo de abandono — M4 Regresión Logística L2 (AUC LOOCV = 0.978, $N = 88$ clientes)

Discusión. M4 cumple el objetivo de predicción del comportamiento comercial al proveer detección proactiva de riesgo de abandono con AUC LOOCV = 0.978. La Regresión Logística supera a XGBoost y Random Forest porque con $N = 88$ la regularización L2 provee mayor estabilidad que los métodos de ensamble, confirmando el principio de parsimonia: con datos limitados, el modelo más simple que captura la señal supera a los más complejos. El modelo habilita intervenciones 90 días antes del abandono detectado, frente al método actual reactivo donde la detección ocurre solo después de que el cliente deja de facturar.

3.2.4 *M5 — Propensión a Expansión*

M5 predice la probabilidad de que un cliente activo cierre al menos una oportunidad ganada en los próximos 6 meses (`expansion_6m`). Dataset B: $N = 88$ clientes, 35 positivos de expansión (39.8 %), 53 negativos (60.2 %). El estudio de ablación con validación LOOCV seleccionó Random Forest (Tabla 3.10):

Algoritmo	AUC LOOCV	AUCPR	Brier	MCC	Seleccionado
RF (200 árboles, prof. 3, balanced)	**0.904**	**0.826**	**0.120**	**0.672**	**Sí**
Regresión Logística L2	0.881	0.775	0.131	0.651	No
XGBoost	0.892	0.873	0.113	0.690	No

Tabla 3.10: Ablación de algoritmos M5 (LOOCV, $N = 88$ clientes)

Las 9 variables activas del modelo M5 (post selección por varianza): mismo conjunto base que M4 más variables adicionales de pipeline y revenue.

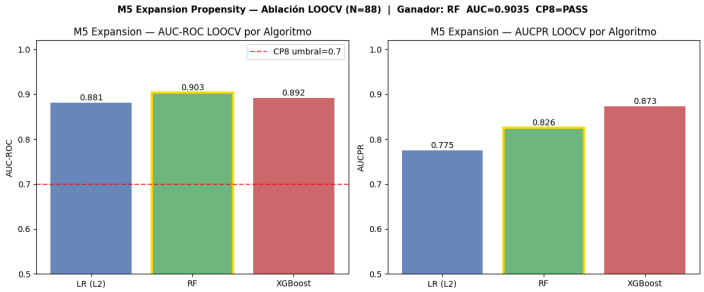

Figura 3.9: Ablación LOOCV y distribución de propensión a expansión — M5 Random Forest (AUC LOOCV = 0.904, $N = 88$ clientes)

Discusión. M5 complementa M4 en el objetivo de predicción del comportamiento comercial al identificar el 39.8 % del portafolio con potencial de expansión en los próximos 6 meses, información que actualmente no existe en forma sistemática en el CRM. Random Forest supera a Regresión Logística en M5 —mientras que LR supera a RF en M4— porque la señal de expansión involucra interacciones no lineales entre `open_opportunities_count`, `won_revenue_total` y `service_breadth_count` que el bosque aleatorio captura sin requerir linealidad. Este contraste confirma que abandono y expansión son fenómenos con estructuras de señal distintas en este portafolio.

3.2.5 *M7 — Predictor de Ciclo de Venta*

M7 predice los días de ciclo de venta sobre $N = 133$ oportunidades ganadas de la cohorte 2025+. El estudio de ablación sobre CV5 seleccionó XGBoost. La Figura 3.10 presenta las predicciones frente a los valores reales:

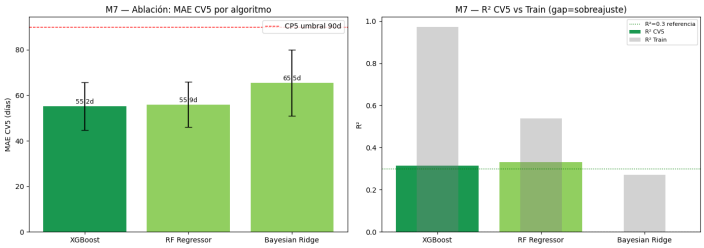

Figura 3.10: Predicciones vs. valores reales — M7 Predictor de Ciclo de Venta (XGBoost, $N = 133$, MAE = 55.2 días, $R^2 = 0.314$)

Las métricas oficiales se presentan en la Tabla 3.11:

Métrica	Valor	Interpretación
Algoritmo	XGBoost	Ganador de ablación vs. RF / Bayesian Ridge
N leads ganados 2025+	133	Umbral $N \geq 75$ superado
R^2 CV5	0.314	Señal real documentada
MAE CV5 (días)	55.2 ± 10.4	Métrica principal (tendencia de portafolio)
Mediana `cycle_days`	76 días	Q1 = 32 d, Q3 = 119 d, máx = 535 d

Tabla 3.11: Métricas de evaluación M7 (Predictor de Ciclo de Venta)

- **Uso recomendado — planificación mensual/trimestral:** MAE = 55 días permite distinguir "oportunidades del próximo mes" frente a "oportunidades del próximo trimestre", sin comprometer fechas específicas.
- **Uso no recomendado — compromisos de fecha individuales:** $R^2 = 0.314$ indica que el 68.6 % de la varianza del ciclo de venta es inherentemente impredecible desde las variables disponibles del CRM (factores externos: decisiones del cliente, procesos de aprobación internos).

Discusión. M7 contribuye al objetivo de métricas de rendimiento en tiempo real al proveer tendencias de ciclo de venta para planificación de portafolio con MAE = 55.2 días sobre una mediana de 76 días. Sus limitaciones son explícitas: $R^2 = 0.314$ indica que el ciclo de ventas en este contexto B2B es inherentemente variable, y el modelo no es apto para SLAs individuales. Su valor operacional es la clasificación del pipeline por percentil de urgencia temporal, no la predicción de fechas exactas de cierre.

3.2.6 *Score Compuesto — Ranking Unificado del Pipeline*

El Score Compuesto integra los outputs de M3, M6 y M7 en una única métrica de priorización para el pipeline activo:

$$\begin{aligned} \text{score\_compuesto} = {} & 0.40 \times p_{\text{cierre}} \\ & + 0.35 \times \frac{\text{salud\_oportunidad}}{100} \\ & + 0.25 \times \text{urgencia\_ciclo} \end{aligned} \tag{3.1}$$

donde `urgencia_ciclo` es el inverso normalizado de los días de ciclo predichos por M7 (mayor urgencia = ciclo más corto = mayor score). Los pesos 0.40/0.35/0.25 priorizan la probabilidad de cierre sobre la salud de la oportunidad y la urgencia temporal.

El umbral operativo > 0.655 produce una tasa de cierre histórica del 93.7 % sobre los $N = 300$ leads que lo superan (11.1 % del pipeline). La distribución del pipeline por zona se presenta en la Tabla 3.12:

Zona	Rango	N leads cerrados	Tasa de cierre
Alto riesgo	< 0.369	—	Baja
Transición	0.369–0.655	—	Moderada
Zona accionable	> 0.655	300 (11.1 %)	**93.7 %**

Tabla 3.12: Distribución del pipeline por zona de Score Compuesto

La Tabla 3.13 traduce cada zona de score en acciones comerciales concretas: define el mensaje diferenciado, el canal de contacto recomendado y la frecuencia de seguimiento para el equipo de ventas.

Rango	Tasa cierre	Perfil	Acción recomendada
> 0.75	$\sim$100 %	Oportunidad madura	Coordinar firmas, preparar incorporación
0.65–0.75	93.7 %	Oportunidad en recta final	Resolver objeciones, acelerar aprobación
0.50–0.65	Moderada	Oportunidad en transición	Recalificar, profundizar necesidades
< 0.35	Baja	Oportunidad en alto riesgo	Análisis de causa raíz, reasignación

Tabla 3.13: Acciones comerciales por rango de Score Compuesto

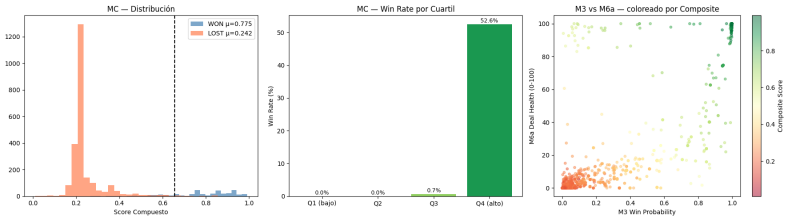

Figura 3.11: Ranking del pipeline por Score Compuesto (brecha WON/LOST = 0.533, $N = 2{,}693$)

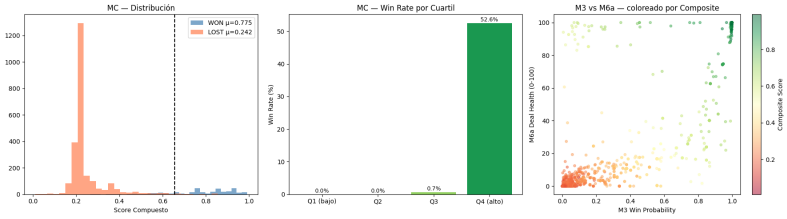

Figura 3.12: Top 20 prospectos por Score Compuesto — lista de priorización accionable

Discusión. El Score Compuesto materializa el objetivo de métricas de rendimiento del equipo comercial en tiempo real al producir una única cifra operacional que integra probabilidad de cierre, salud de la oportunidad y urgencia temporal. El umbral de 0.655 selecciona el 11.1 % del pipeline con tasa de cierre histórica del 93.7 %, permitiendo al equipo concentrar su esfuerzo en las oportunidades con mayor retorno esperado sin criterio de clasificación subjetivo. La separación WON/LOST de 0.533 es resultado directo de tres decisiones de diseño: M3 sobre $N = 2{,}693$ con XGBoost (AUC CV5 $= 0.971$), M6 arquitecturalmente independiente (Spearman $= 0.718$) y M7 sobre la cohorte 2025+ (MAE $= 55.2$ días).

3.3 *Importancia de Variables e Interpretabilidad del Pipeline*

El análisis de importancia de variables permite identificar qué información del CRM determina cada predicción del pipeline, ofreciendo una capa de interpretabilidad que complementa las métricas de rendimiento reportadas por modelo.

M3 — Win Probability (Dataset A). El análisis SHAP sobre una muestra estratificada de leads identifica la urgencia temporal de la oportunidad —medida por los días disponibles hasta la fecha límite— como el predictor de mayor peso, con una contribución muy superior al resto de variables. El valor esperado de la oportunidad ocupa el segundo lugar, mientras que un conjunto de variables de temporalidad de apertura —mes, trimestre y día de la semana de registro— forma un tercer grupo de señal estacional. La prioridad registrada por el ejecutivo de cuenta y el nivel de actividad comunicacional completan el perfil de variables activas. La Figura 3.13 muestra la distribución de contribuciones SHAP sobre la muestra de evaluación.

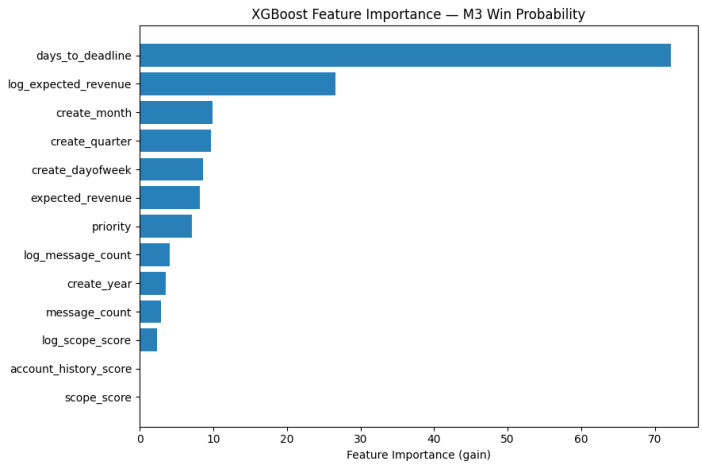

Figura 3.13: Análisis SHAP — M3 Win Probability (contribuciones por variable sobre muestra estratificada de leads)

M6 — Deal Health Score (Dataset A). M6 opera sobre el mismo conjunto de variables predictoras que M3, pero con un objetivo de regresión continua que cuantifica la madurez de la oportunidad. La importancia por ganancia muestra un perfil distinto: las variables de alcance del proyecto y de relacionamiento con el prospecto ganan peso relativo frente a la urgencia temporal. Esto refleja que la madurez de una oportunidad depende más del perfil acumulado del prospecto que de su posición en el calendario, a diferencia de la probabilidad de cierre. La Figura 3.14 ilustra la distribución de importancia para este modelo.

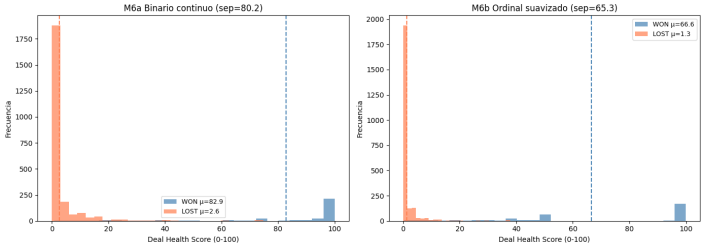

Figura 3.14: Importancia de variables por ganancia — M6 Deal Health Score (XGBRegressor, Dataset A)

M7 — Predictor de Ciclo de Venta (Dataset A). La urgencia temporal vuelve a encabezar la importancia de variables en M7, confirmando que el horizonte de la oportunidad es el factor estructural dominante en todos los modelos del Dataset A. La diferencia notable respecto a M3 y M6 es la aparición del score PESA agregado entre las variables de mayor contribución: M7 es el único modelo del Dataset A donde el perfil de relacionamiento del prospecto aporta señal directa al output, en este caso para estimar la duración esperada del ciclo de venta.

M4 y M5 — Riesgo de Abandono y Propensión a Expansión (Dataset B). En el Dataset B, las variables activas pertenecen a un universo distinto: historial de facturación, relacionamiento longitudinal y portafolio de servicios. El score PESA agregado y el indicador de salud de cuenta son las variables de mayor peso en ambos modelos, confirmando que el perfil relacional del cliente —no la urgencia puntual— determina el riesgo de abandono y la propensión a expansión. La recencia de facturación aparece exclusivamente en M5 y no en M4, señalando que la distancia desde la última transacción es predictora de oportunidad de expansión pero no de abandono.

Patrón transversal. Los modelos del Dataset A responden principalmente a la dinámica temporal del ciclo de venta y al valor económico de la oportunidad; los modelos del Dataset B responden al historial relacional acumulado del cliente a través de las dimensiones del framework PESA. Esta división confirma el diseño bidimensional del sistema: los modelos de prospección capturan la urgencia del pipeline activo, mientras que los modelos de retención capturan la

profundidad de la relación comercial establecida.

4 Conclusiones y Trabajo Futuro

4.1 Conclusiones

Las conclusiones se organizan en torno a los tres componentes del objetivo general, verificando qué se logró en cada uno y bajo qué condiciones.

4.1.1 Segmentación dinámica de prospectos y clientes

El sistema produce clasificación dinámica y automática de prospectos y clientes en grupos con comportamiento comercial diferenciado. Para los prospectos, la frontera más determinante es la existencia de una cotización formal: su presencia o ausencia separa perfiles con tasas de cierre radicalmente distintas, evidenciando que el framework PESA captura estructura real en los datos operativos del CRM. Para los clientes activos, el agrupamiento distingue perfiles de riesgo de abandono y potencial de expansión que no existían en forma sistemática, permitiendo identificar de manera directa los segmentos que requieren intervención y los que representan oportunidad activa de crecimiento. La convergencia entre el agrupamiento por comportamiento y la detección por perfil financiero atípico —dos métodos independientes que señalan la misma estructura— refuerza la robustez de los resultados como evidencia complementaria de la misma señal subyacente en el pipeline.

Verificación del objetivo general — Segmentación dinámica. El objetivo de agrupamiento dinámico de prospectos y clientes se cumple: el sistema produce perfiles empíricamente validados que no existían en el CRM antes de este trabajo. La clasificación automática permite concentrar el esfuerzo comercial en los segmentos de mayor rendimiento probable sin depender del criterio subjetivo del vendedor, y provee una base estructurada para la incorporación futura de herramientas de inteligencia generativa que operen en contexto del perfil de cada prospecto o cliente.

4.1.2 *Predicciones de comportamiento comercial*

Los cinco modelos supervisados proveen señales predictivas cuantificadas donde antes solo existía el juicio empírico del equipo comercial. La estimación de cierre registrada en el CRM por los vendedores no tiene valor discriminativo —como se documenta en el capítulo de resultados—, por lo que el sistema representa una mejora absoluta respecto al punto de partida. El aporte más inmediato es la habilitación de detección proactiva: el riesgo de abandono se identifica con antelación suficiente para actuar, el potencial de expansión del portafolio se cuantifica de forma sistemática, y la salud y urgencia de cada oportunidad activa dispone de una señal objetiva de referencia para el equipo comercial.

Un hallazgo metodológico relevante es que fenómenos del mismo portafolio —abandono y expansión— responden a estructuras de señal distintas: el primero es capturable con modelos lineales parsimoniosos, el segundo requiere capturar interacciones entre variables que los modelos de ensamble manejan mejor. Esto confirma el valor de evaluar múltiples algoritmos para cada problema en lugar de asumir una solución única.

Verificación del objetivo general — Predicciones de comportamiento comercial. El objetivo de predicción de comportamiento comercial se cumple: los modelos supervisados superan en todos los casos los criterios de aceptación definidos en §2.4 y proveen señales analíticas reproducibles y actualizables que no existían antes de este sistema.

4.1.3 *Métricas de rendimiento en tiempo real*

El tercer componente se materializa en el Score Compuesto y el registro de modelos. El Score Compuesto integra las salidas de los tres modelos de prospecto en una única cifra operacional que permite ordenar el pipeline por prioridad de cierre sin necesidad de interpretar múltiples indicadores por separado. Antes de este sistema, no existían métricas estandarizadas de rendimiento comercial en la organización: la única referencia disponible era la estimación subjetiva del vendedor, sin valor discriminativo estadístico.

Verificación del objetivo general — Métricas de rendimiento en tiempo real. El objetivo de métricas de rendimiento en tiempo real se cumple: el sistema provee una puntuación operacional única con umbral documentado, criterio de reentrenamiento y registro de artefactos que garantiza replicabilidad y mantenimiento en producción. La introducción de cualquier señal cuantitativa validada representa una mejora absoluta respecto a la ausencia de sistema que existía antes de este trabajo.

4.1.4 Limitaciones del estudio

Las limitaciones del estudio se agrupan en tres categorías.

Volumen de datos. Los modelos de comportamiento de clientes operan sobre un número reducido de observaciones, insuficiente para validación cruzada estándar; se optó por una estrategia de validación que entrena con todos los casos menos uno, lo que provee una estimación insesgada pero con mayor incertidumbre que la obtenida con conjuntos más grandes. El predictor de ciclo de venta opera exclusivamente sobre las oportunidades ganadas de la cohorte más reciente, dado que los registros históricos presentan fechas de cierre imputadas de forma retroactiva que distorsionarían el entrenamiento. La segmentación de clientes activos refleja la heterogeneidad real del portafolio: la alta dispersión en el valor facturado hace que ninguna partición produzca grupos perfectamente homogéneos.

Especificidad temporal y sectorial. El conjunto de datos captura un único contexto empresarial —Soltein, empresa partner de Odoo en México, período 2022–2025— lo que limita la generalización directa del framework. La imputación retroactiva de fechas de cierre en la carga inicial de datos históricos introduce valores anómalos en la duración del ciclo de venta que el modelo trata como variabilidad estocástica, pero que podrían esconder un sesgo sistemático. La generalización del framework PESA y los umbrales de los perfiles de cliente a otros contextos de consultoría ERP requiere revalidación empírica.

Limitaciones metodológicas. El vendedor asignado a cada oportunidad genera dependencia estadística entre leads del mismo ejecutivo de cuenta: los estilos individuales de registro y seguimiento producen patrones sistemáticos que el modelo no aisló explícitamente. La diferencia entre el rendimiento observado en validación cruzada y el observado sobre datos de períodos posteriores cuantifica el costo del cambio natural en el comportamiento del pipeline a lo largo del tiempo, y representa la limitación operacional más relevante para el despliegue y mantenimiento del sistema en producción.

4.2 Trabajo Futuro

4.2.1 Modelos habilitados y diferidos

Tres modelos de agregación analítica están diseñados y diferidos por insuficiencia de datos en el momento del estudio: pronóstico de revenue, calidad de fuente de prospectos y desempeño por vendedor. El primero requiere al menos doce meses de datos de producción para construir series temporales confiables; los dos últimos requieren corrección en el campo de equipo de ventas, que actualmente concentra la mayoría de los registros en un único responsable. Dos modelos adicionales

—monitor de riesgo de contratos recurrentes y valor de vida del cliente— quedan condicionados a alcanzar el volumen mínimo de observaciones necesario para entrenamiento confiable.

4.2.2 *Integración de inteligencia generativa: variables cualitativas LLM*

El framework PESA captura señales operativas del CRM, pero no puede derivar automáticamente señales cualitativas como la autoridad del decisor, el presupuesto disponible, la urgencia declarada por el prospecto o el sentimiento del historial de comunicaciones. Estas señales existen en el historial de correos y notas del CRM pero requieren inferencia mediante modelos de lenguaje de gran escala para ser incorporadas como variables predictoras. Su integración ampliaría el framework PESA con dimensiones cualitativas adicionales, con potencial de mejora en los modelos supervisados de clasificación y salud de oportunidades.

4.2.3 *MLOps y reentrenamiento automático*

Para que el sistema sea sostenible en producción se requieren cuatro componentes: reentrenamiento periódico de los modelos de clasificación y salud de oportunidades; monitoreo automático de cambios en el comportamiento de los datos de entrada, con alerta cuando los valores observados se alejan del perfil histórico de entrenamiento; registro de versiones de modelos con capacidad de revertir a una versión anterior si el nuevo modelo degrada el rendimiento; y notificaciones automáticas al equipo responsable cuando el sistema detecte condiciones anómalas en el pipeline de datos.

4.2.4 *Validación de coherencia entre modelos*

Una línea de validación pendiente consiste en verificar que los modelos del sistema producen señales coherentes entre sí cuando se aplican sobre el mismo prospecto o cliente. En particular, se espera que los perfiles de mayor valor identificados por el agrupamiento de prospectos presenten también las probabilidades de cierre más altas, y que las oportunidades con mayor puntuación de salud se encuentren en etapas activas del pipeline. Validar estas relaciones de forma sistemática, y definir alertas cuando se detecten inconsistencias, garantizaría la integridad del sistema ante cambios en los datos de entrada o en los procesos de extracción.

4.2.5 *Extensión a otros contextos de consultoría*

El framework PESA y los algoritmos implementados son transferibles a otros contextos de consultoría ERP con ajuste de pesos y umbrales de arquetipos. Una línea de investigación futura es validar el framework en otras empresas partner de Odoo con perfiles similares (consultoría ERP, ciclos largos, servicios recurrentes), lo que permitiría establecer puntos de referencia sectoriales y evaluar la generalización del sistema.

Reflexión Final

Este trabajo ha demostrado que la inteligencia comercial moderna no requiere infraestructura de grandes volúmenes de datos ni repositorios centralizados. Los ERP comerciales como Odoo proporcionan datos operativos suficientemente ricos para entrenar modelos que generan valor inmediato en la toma de decisiones.

El verdadero desafío en inteligencia comercial radica en formular las preguntas correctas: ¿cuáles son las dimensiones que verdaderamente importan para agrupar prospectos y clientes en contextos específicos? Una vez respondida esa pregunta con rigor metodológico, los datos y los algoritmos se convierten en herramientas para amplificar la capacidad de decisión del equipo comercial.

Customer DNA proporciona una respuesta a esa pregunta para empresas de consultoría de ERP. Esperamos que otros equipos comerciales e investigadores se beneficien del framework y continúen explorando la intersección entre estrategia comercial y aprendizaje automático.

Bibliografía

[1] Salesforce, "State of sales, 6th edition." https://www.salesforce.com/resources/research-reports/state-of-sales/, 2024. Salesforce Research.

[2] A. M. Hughes, *Strategic Database Marketing*. Chicago: Probus Publishing, 1994.

[3] M. Bohanec, M. Kljajić Borštnar, and M. Robnik-Šikonja, "Explaining machine learning models in sales predictions," *Expert Systems with Applications*, vol. 71, pp. 416–428, Apr. 2017.

[4] J. R. Bult and T. Wansbeek, "Optimal selection for direct mail," *Marketing Science*, vol. 14, no. 4, pp. 378–394, 1995.

[5] T. Tran, T. Nguyen, and H. Le, "Customer churn prediction using machine learning in B2B SaaS: A comparative study," *IEEE Access*, vol. 11, pp. 34120–34133, 2023.

[6] D. Kahneman and D. Lovallo, "Timid choices and bold forecasts: A cognitive perspective on risk taking," *Management Science*, vol. 39, pp. 17–31, Jan. 1993.

[7] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, (Berkeley, CA, USA), pp. 281–297, University of California Press, 1967.

[8] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, NY, USA: Wiley, 1990.

[9] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE International Conference on Data Mining (ICDM)*, (Pisa, Italy), pp. 413–422, 2008.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY, USA), pp. 785–794, ACM, 2016.

[11] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society, Series B*, vol. 20, no. 2, pp. 215–242, 1958.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[14] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[16] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York: Wiley, 2nd ed., 2000.

[17] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.