Problema Empresarial:

Una institución bancaria portuguesa enfrenta ineficiencias significativas en sus campañas de marketing directo telefónico para depósitos a plazo. Según datos históricos (2008-2010), se requieren múltiples contactos por cliente para determinar su interés, lo que genera costos operativos elevados y una baja productividad del equipo de telemarketing. La ausencia de una segmentación predictiva resulta en contactos masivos e indiscriminados, desperdiciando recursos en prospectos con baja probabilidad de conversión. Esto impacta directamente al departamento de marketing (ROI deficiente), al **centro de llamadas (call center)** (productividad reducida) y a la gestión comercial (metas de captación no alcanzadas), afectando así la estrategia de crecimiento de la cartera de productos de inversión del banco.

· Justificación del Uso de Ciencia de Datos e IA:

Mediante el **Aprendizaje Automático (Machine Learning)**, podemos implementar un modelo predictivo que mejore la eficiencia, analizando los patrones en los datos categóricos de este **conjunto de datos (dataset)**. Paralelamente, utilizando modelos como la **regresión logística (logistic regression)**, concepto que estamos estudiando en el curso de **aprendizaje automático (machine learning)**, podríamos desarrollar un modelo para clasificar si un cliente contratará o no un depósito a plazo.

· Formulación de Pregunta SMART:

¿Es posible crear un modelo de **Aprendizaje Automático (Machine Learning)**, utilizando el conjunto de datos bank-additional-full.csv, que prediga la suscripción a un depósito a plazo con una **exactitud (accuracy)** mínima de entre el 70 % y el 85 %? El objetivo es aplicar este modelo a un conjunto de 500 prospectos para alcanzar una tasa de conversión del 10 % al 30 %.

- **S (Específico - Specific):** Predecir qué clientes de un nuevo grupo de 500 prospectos se suscribirán a un depósito a plazo, identificando a aquellos con mayor probabilidad para optimizar el contacto.
- **M (Medible - Measurable):** Lograr una **exactitud (accuracy)** global del modelo de al menos el 85 % y una **precisión (precision)** de al menos el 70 % para la clase positiva (clientes que sí se suscriben). Se busca obtener una tasa de conversión del 10 % al 30 % en el grupo de prospectos seleccionados por el modelo.
- **A (Alcanzable - Achievable):** Con el conjunto de datos proporcionado, es factible alcanzar el rendimiento esperado del modelo.
- **R (Relevante - Relevant):** Este modelo puede tener un impacto significativo en las metas de la organización, aumentando la eficiencia del **centro de llamadas (call center)** y el retorno de la inversión.
- **T (Temporal - Time-bound):** El objetivo debe alcanzarse antes de que finalice el presente semestre académico.

· Justificación técnica:

El contenido de las variables predictoras y el volumen de datos deberían permitir desarrollar los objetivos del proyecto de este curso en un tiempo razonable y con métricas medibles.

Referencias.

- Yamahata, H. (n.d.). *Bank Marketing* [Data set]. Kaggle. Recuperado el 16 de agosto de 2025 de https://www.kaggle.com/datasets/henriqueyamahata/bank-marketing. (Kaggle)
- Moro, S., Rita, P., & Cortez, P. (2014). *Bank Marketing* [Data set]. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306. (UCI Machine Learning Repository)
- Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems, 62*, 22–31. https://doi.org/10.1016/j.dss.2014.03.001. (scirp.org)

Actualización de la pregunta Smart

A nivel Mediable-Mesurable:

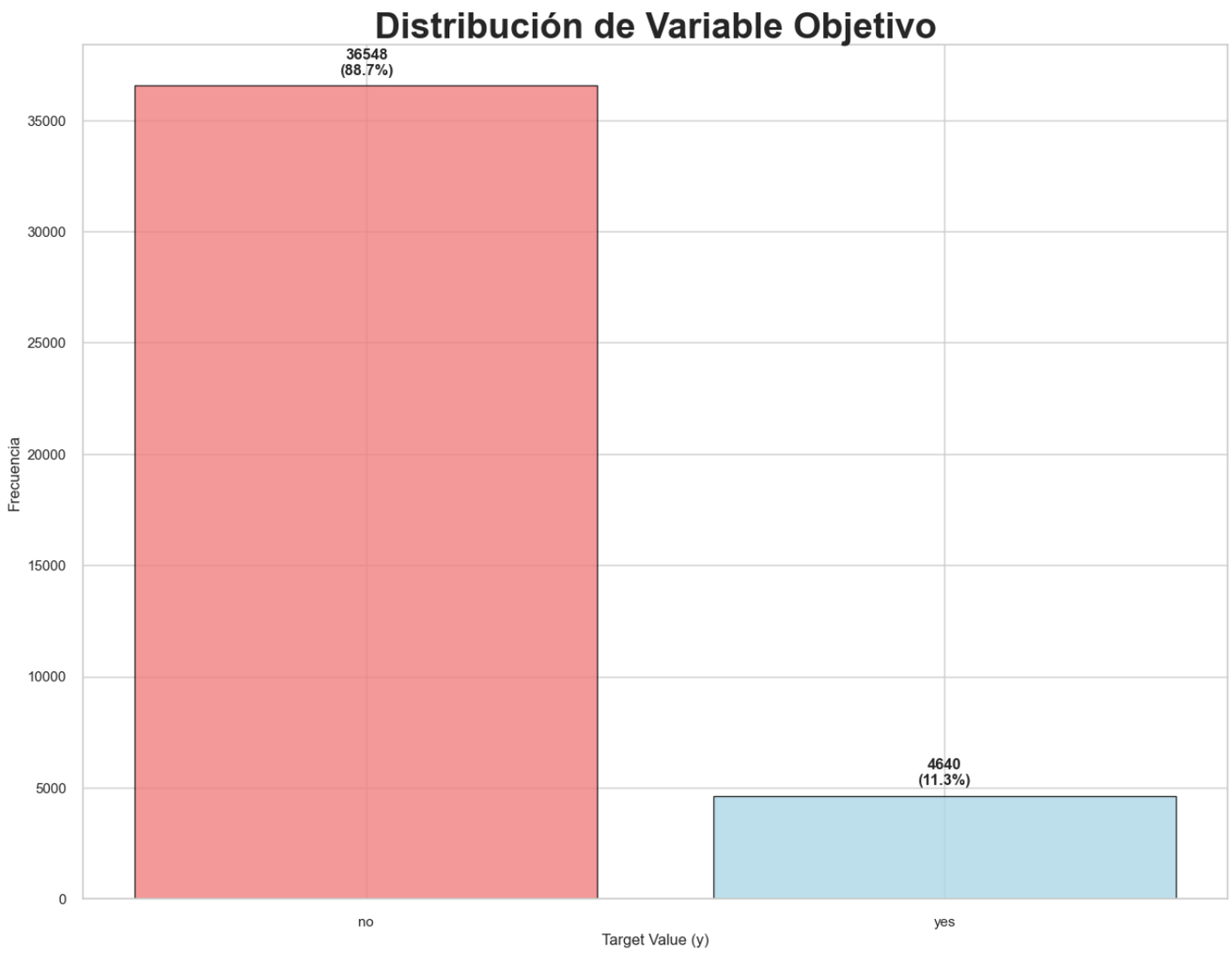
La métrica debería ser redefinida en términos de la manera mediante las opciones presentadas a continuación.

- **Opción 1.**
  - **Desarrollar un modelo que permita contactar al top 20% de prospectos con mayor probabilidad, logrando una tasa de conversión del 15-25% en este segmento (vs. la tasa base del datase).**
- **Opción 2**
  - **Lograr un recall mínimo del 60% para la clase positiva y una precision de al menos 35%, con el objetivo de no perder más del 40% de clientes potenciales.**

Estrategia de Manejo de Valores Faltantes

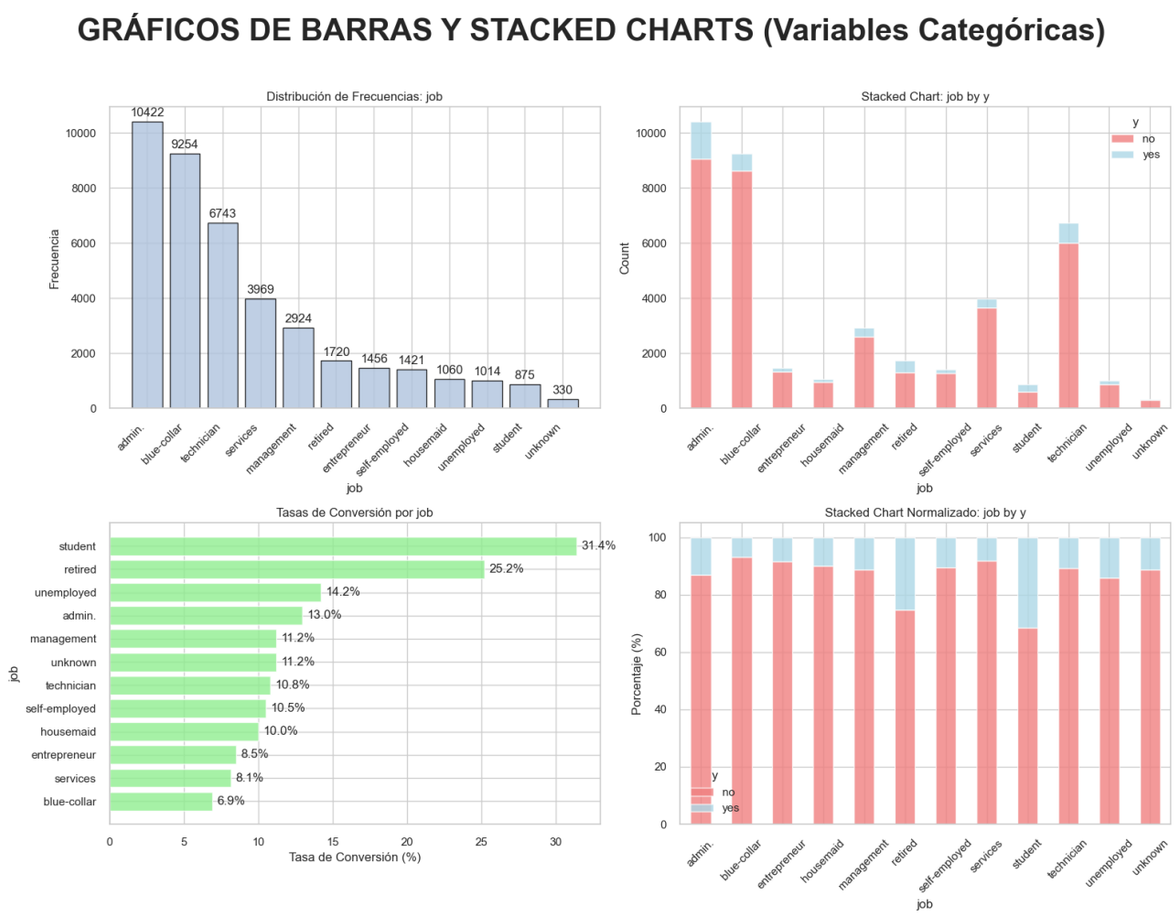
Se tomo una estrategia para el tratamiento de valores faltantes, manteniendo las categorías "unknown" como información válida en lugar de realizar imputación o eliminación de registros. Esta decisión se soporta en lo siguiente:

- **Los valores "unknown" pueden contener patrones de comportamiento específicos y valiosos**
- **Preservan la naturaleza original de los datos sin introducir sesgos artificiales**
- **Representan un segmento real de clientes con información incompleta**



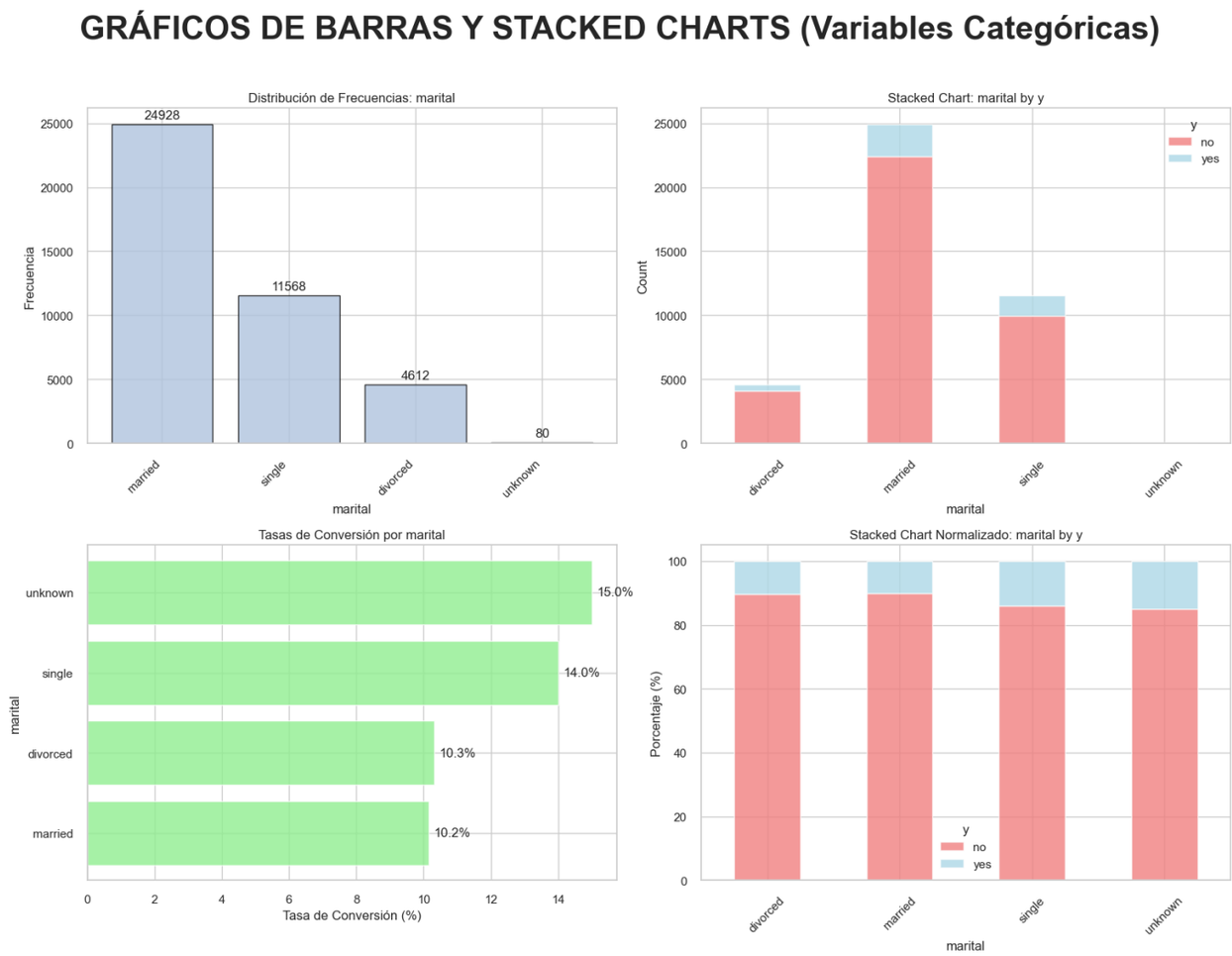
HALLAZGOS DEL ANÁLISIS EXPLORATORIO

- **Grupos etarios favorables: Clientes mayores, jubilados y estudiantes muestran mayor probabilidad de conversión**
- **Perfiles ocupacionales: Trabajadores en sectores de servicios y administración presentan tasas de respuesta diferenciadas**



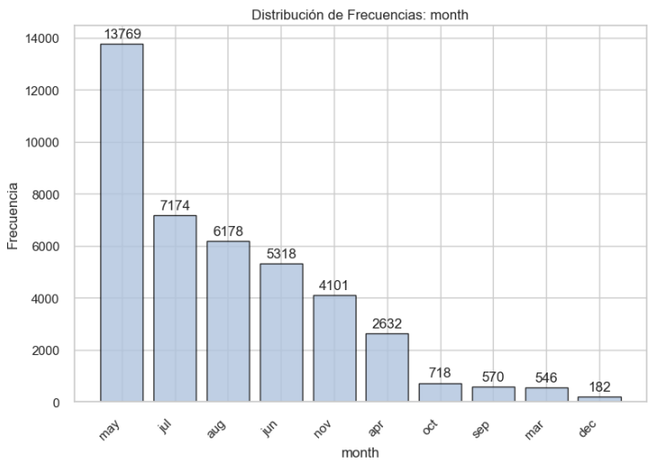
-

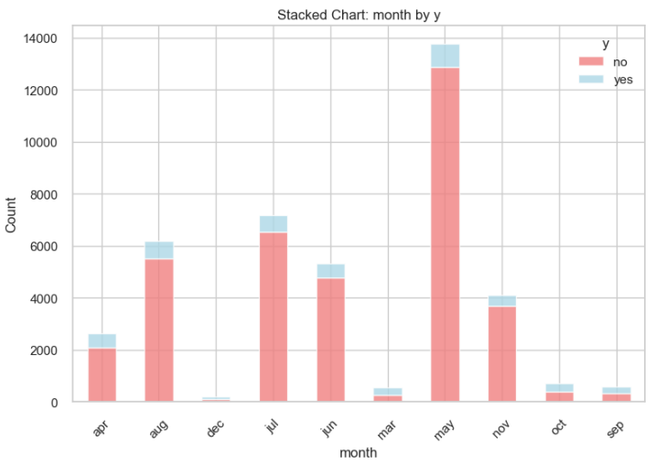
- **Estado civil: Se identificaron patrones específicos relacionados con el estado civil y la decisión de inversión**

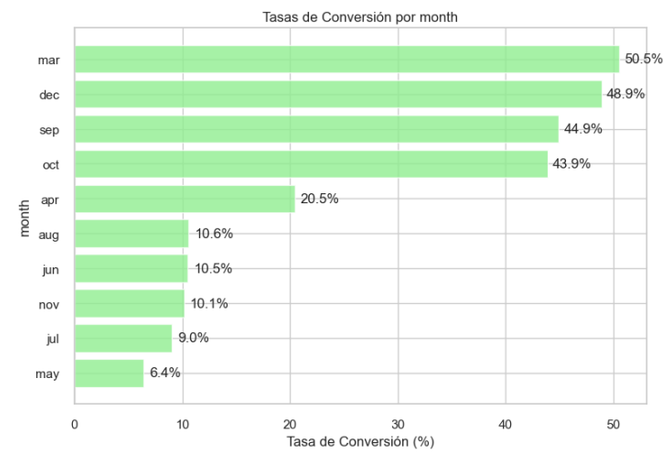


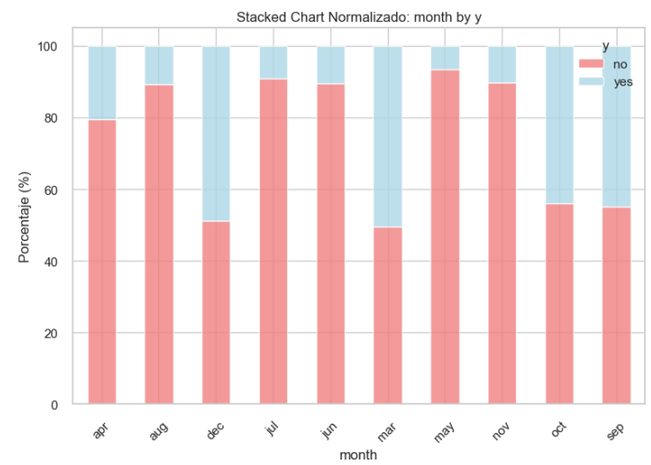
- **Patrón Estacional Crítico:**
  - Los meses de marzo, septiembre, octubre y diciembre presentan tasas de conversión significativamente superiores al promedio, sugiriendo momentos óptimos para intensificar las campañas de marketing.

GRÁFICOS DE BARRAS Y STACKED CHARTS (Variables Categóricas)



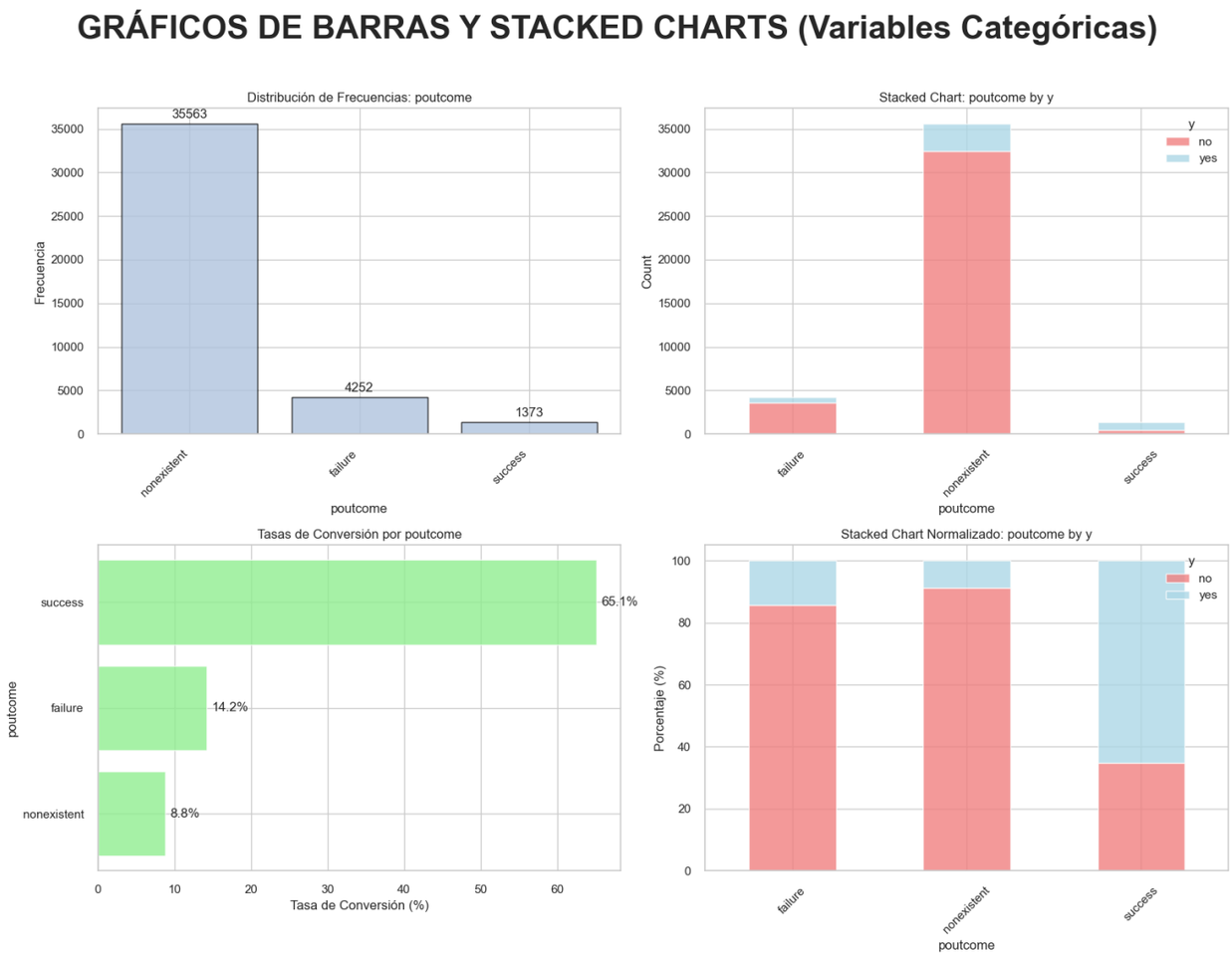






Impacto del Historial de Contacto

Se identificó una correlación fuerte entre el historial de campañas anteriores y la probabilidad de suscripción actual. Los clientes con campañas anteriores exitosas tienen una probabilidad significativamente mayor de suscribir nuevos productos.

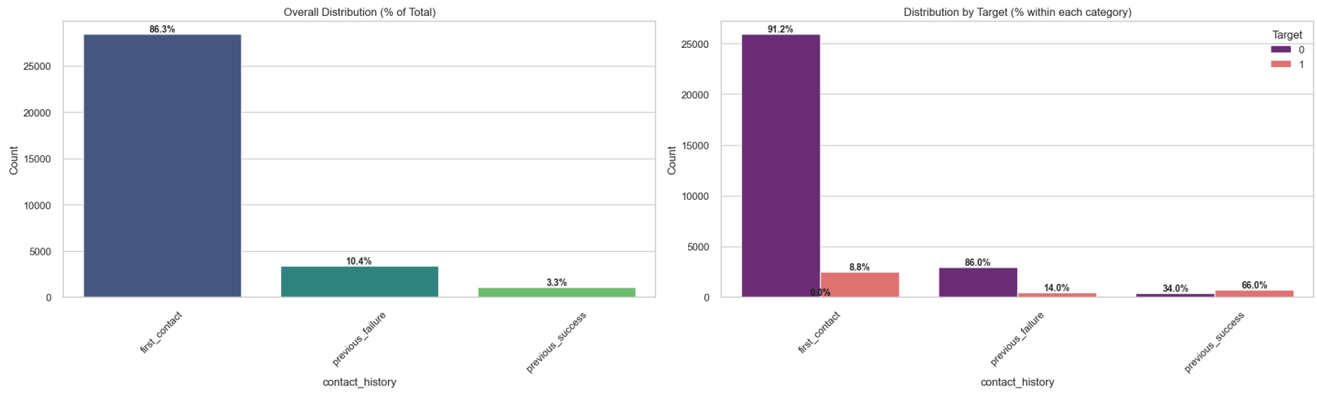


DECISIONES DE INGENIERÍA DE CARACTERÍSTICAS

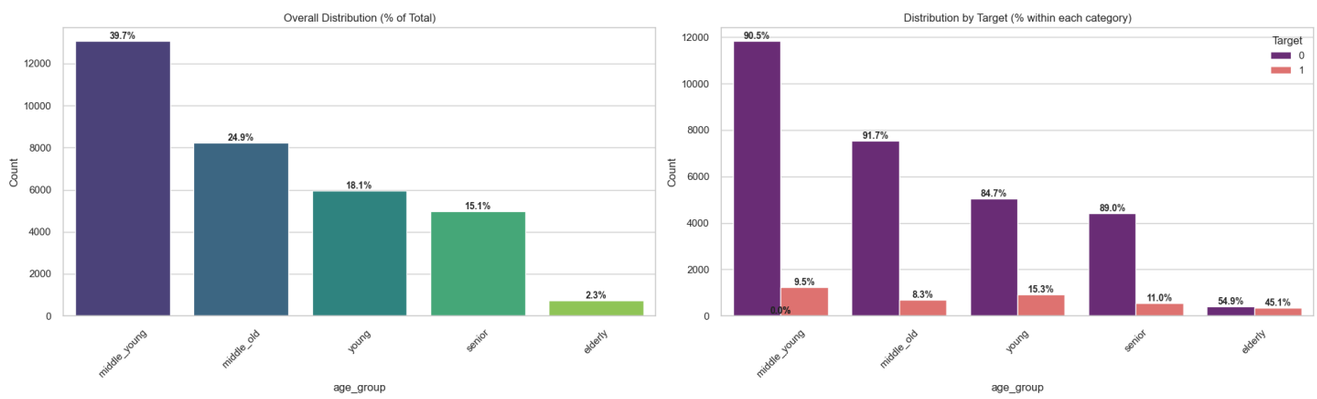
La creación de cuatro nuevas variables derivadas para capturar patrones no lineales importantes:

Variable Nueva	Propósito	Impacto
age_group	Segmentación etaria	Captura patrones no lineales de edad. Se calcula en función de la columna edad
campaign_intensity	Intensidad de contacto	Optimiza frecuencia de llamadas. Se calcula a partir de la columna campaing
contact_history	Historial de interacción	Mejora predicción basada en experiencia previa. Se calcula en función de la columna poutcome
economic_context	Contexto macroeconómico	Simplifica variables económicas correlacionadas. Se calcula en función de nr.employed

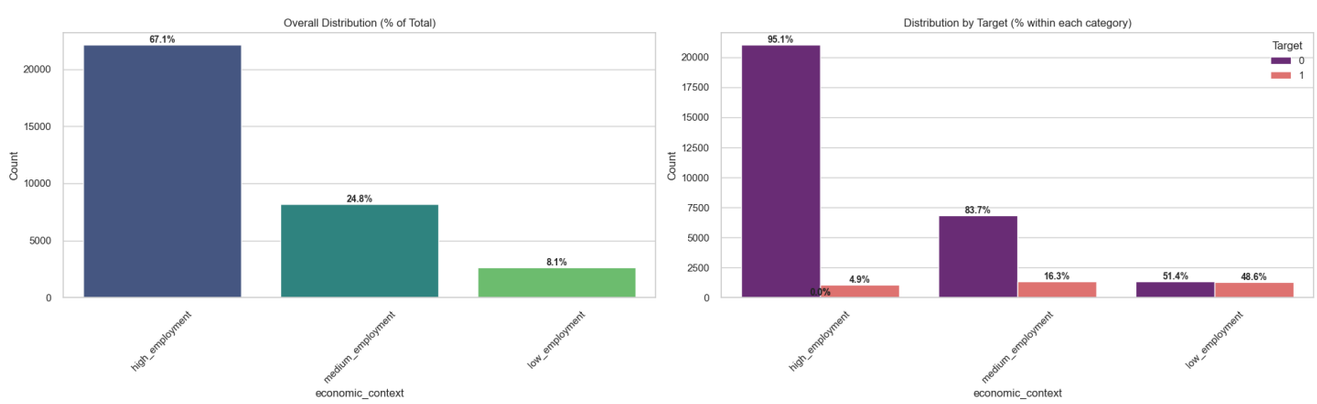
Analysis of Categorical Column: 'contact_history'



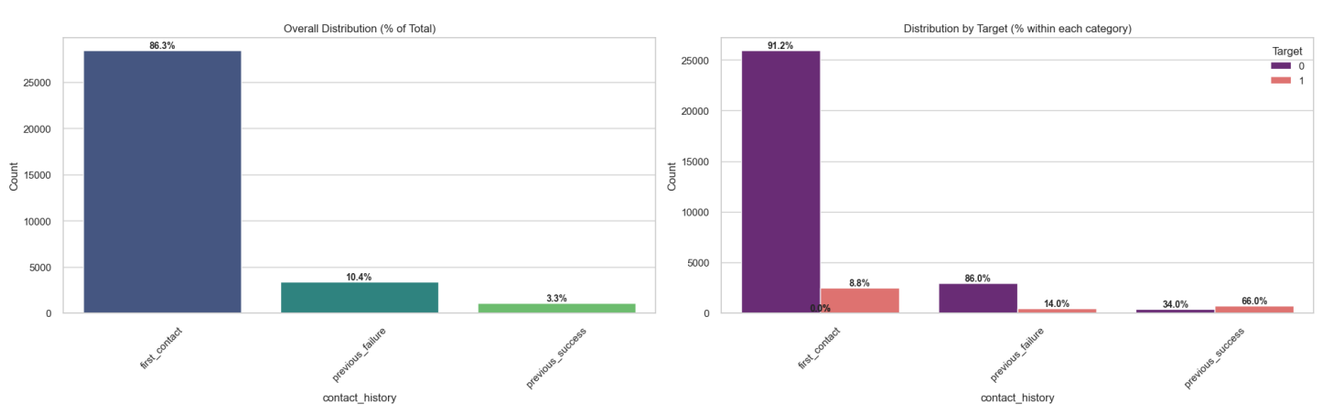
Analysis of Categorical Column: 'age_group'



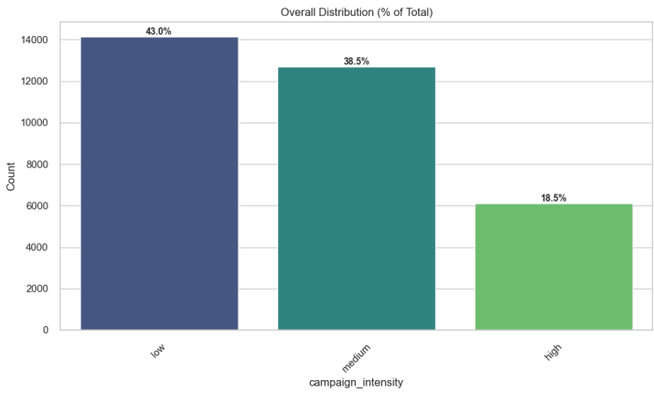
Analysis of Categorical Column: 'economic_context'



Analysis of Categorical Column: 'contact_history'



Analysis of Categorical Column: 'campaign_intensity'

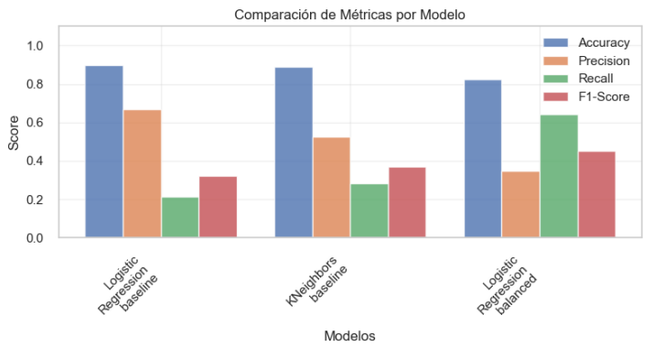


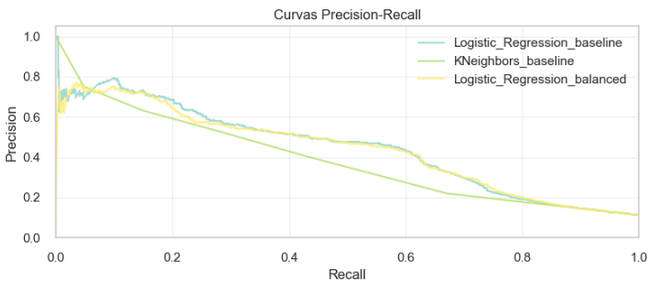
Resultados iniciales del modelo de clasificación.

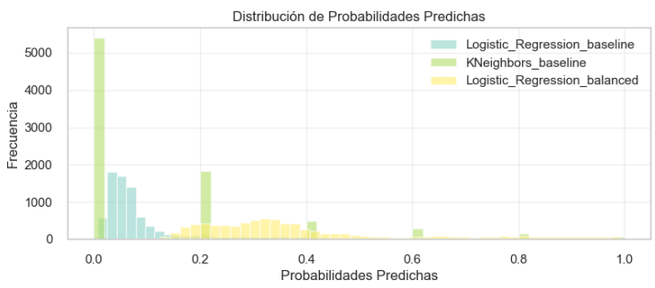
- **Regresion logistica simple:**
  - **Accuracy: 0.8993**
  - **Precision: 0.6667**
  - **Recall: 0.2134**
  - **F1-Score: 0.3233**

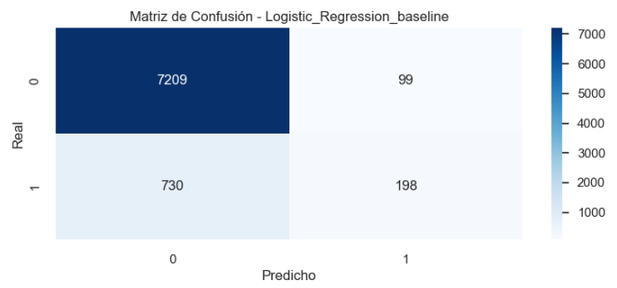
- **KNN**
  - **Accuracy: 0.8904**
  - **Precision: 0.5249**
  - **Recall: 0.2845**
  - **F1-Score: 0.369**

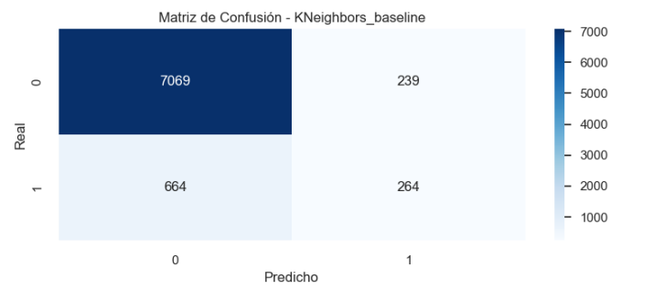
- **Regresion logistica balanceada:**
  - **Accuracy: 0.8248**
  - **Precision: 0.3492**
  - **Recall: 0.6422**
  - **F1-Score: 0.4524**

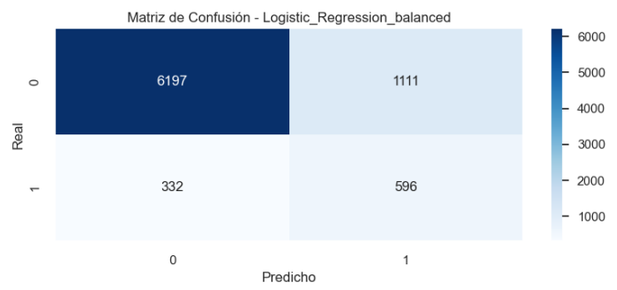












Conclusión:

Basado en los modelos presentados, ninguno logra una F1 de más del 50%, por lo tanto, se debe iterar. En paralelo, la regresión logística balanceada es la que mejor se adapta a la forma de los datos debido a un dataset desbalanceado, gracias a que este modelo presta atención a los datos de la clase minoritaria para aprender mejor el patrón de los mismos.