Universidad del Desarrollo

Machine Learning

Dataset: Marketing Bancario Bank-Full (UCI / Kaggle)

Integrantes

Cristian Vargas

Claudio Ballerini

Nector Felipe Ruiz

Diciembre,2025

1 Introducción

Las instituciones financieras enfrentan el desafío de optimizar sus campañas comerciales, especialmente aquellas relacionadas con productos de baja frecuencia de contratación, como los depósitos a plazo. Tradicionalmente, el Banco de Portugal realizaba campañas telefónicas masivas dirigidas a toda su base de clientes, lo que implica altos costos operacionales, baja eficiencia y una tasa de conversión limitada debido al escaso nivel de personalización.
El problema central consiste en predecir qué clientes tienen una mayor probabilidad de aceptar un depósito a plazo, para así orientar las campañas hacia segmentos más receptivos y maximizar el retorno de inversión.
Para abordar este problema se utiliza el Bank Marketing Dataset, publicado originalmente por Moro et al. (UCI Machine Learning Repository). Este dataset contiene más de 45.000 registros de campañas telefónicas reales e incluye variables demográficas, socioeconómicas y operacionales, tales como:

Información del cliente (edad, profesión, estado civil, educación)
Condiciones financieras individuales (saldo, préstamos)
Variables relacionadas con la campaña (mes, número de contactos previos, duración de la llamada)
Indicadores macroeconómicos y de contexto
La variable objetivo 'y', que indica si el cliente contrató un depósito a plazo
Este dataset es adecuado por varias razones:
Refleja un caso real de negocio, con resultados observados en campañas históricas.
Contiene una mezcla de variables numéricas y categóricas, ideal para modelos avanzados como XGBoost.
Su gran tamaño permite aplicar técnicas robustas de machine learning.
Es un benchmark ampliamente utilizado para estudios de marketing predictivo.
En consecuencia, el dataset es idóneo para construir un pipeline completo de análisis predictivo, aplicar técnicas supervisadas y no supervisadas, evaluar modelos y generar recomendaciones accionables para campañas comerciales.

2 Justificación Dataset y EDA

Según el análisis del dataset Esta es una propuesta formal y estructurada para la justificación del Análisis Exploratorio de Datos (EDA) de tu dataset "Bank Marketing".

Puedes usar esta estructura para tu informe o presentación, ya que aborda la relevancia del dataset desde una perspectiva de negocio, técnica y estratégica.

Justificación del EDA: Bank Marketing Dataset
El Análisis Exploratorio de Datos (EDA) sobre el conjunto de datos de marketing bancario es fundamental y se justifica bajo cuatro pilares estratégicos: la optimización del retorno de inversión (ROI), la segmentación de clientes, la eficiencia operativa de las campañas y la calidad de los datos para modelado predictivo.

1. Relevancia de Negocio y Optimización de Recursos

El objetivo principal (variable y) es predecir si un cliente suscribirá un depósito a plazo. Las campañas de telemarketing son costosas en términos de tiempo humano y recursos técnicos.

Un EDA permite identificar patrones que distinguen a los clientes que compran (yes) de los que no (no). Entender esto permite a la institución bancaria dirigir sus esfuerzos solo a los clientes con alta probabilidad de conversión, reduciendo costos operativos y aumentando la tasa de éxito.

2. Riqueza Dimensional para la Segmentación (Profiling)

El diccionario de datos presenta una combinación robusta de atributos que permite crear perfiles de clientes detallados.

Demográficos (age, job, marital, education): El EDA revelará qué grupos demográficos son más propensos al ahorro. Ejemplo: ¿Tienen los jubilados (retired) o los estudiantes (student) una mayor tasa de suscripción que los trabajadores manuales (blue-collar)?

Financieros (balance, housing, loan, default): Permite analizar la salud financiera del cliente. Se justifica explorar si tener deudas vigentes (housing, loan) correlaciona negativamente con la capacidad de abrir un nuevo depósito.

3 Análisis de la Dinámica de la Campaña y Estacionalidad

Las variables relacionadas con el contacto ofrecen insights operativos críticos que no dependen del perfil del cliente, sino de la ejecución de la estrategia.

Fatiga del cliente (campaign, previous): Es vital analizar si existe un punto de inflexión donde aumentar el número de llamadas (campaign) comienza a ser contraproducente y genera rechazo.

Estacionalidad (day, month): El EDA justificará si existen meses específicos (ej. mayo vs. diciembre) donde la predisposición al ahorro aumenta, permitiendo planificar futuras campañas en ventanas de tiempo óptimas.

Impacto del historial (poutcome, pdays): Analizar el éxito de campañas previas es el mejor predictor del comportamiento futuro.

4 Calidad de Datos y Preparación para Modelado (Machine Learning)

Antes de aplicar cualquier algoritmo predictivo, el EDA es indispensable para asegurar la integridad técnica:

Variable duration: El diccionario advierte que esta variable afecta fuertemente al target. El EDA debe confirmar esto y justificar su exclusión si el objetivo es un modelo predictivo realista (ya que la duración no se conoce antes de hacer la llamada), o su inclusión si es solo para análisis descriptivo post-mortem.

Valores Atípicos y Nulos: Variables como balance (con rangos de -8,000 a 100,000+) y pdays (con valor -1) requieren un análisis de distribución para decidir estrategias de limpieza o transformación.

Desbalance de Clases: Es altamente probable que la variable y esté desbalanceada (muchos 'no', pocos 'yes'). El EDA cuantificará este desbalance para justificar técnicas posteriores de remuestreo (SMOTE, Undersampling).

Tiene una distribución de la siguiente manera:

5 Interpretación de segmentos K-Means

La imagen generada en nuestro notebook (el mapa de calor) nos muestra cómo se relacionan las variables numéricas entre sí: Tras evaluar la estructura de los datos con y hemos seleccionado como la configuración óptima para el negocio.
A continuación, se describen los perfiles identificados y la estrategia sugerida para cada uno:

Clúster 0: "Prospectos de Alto Potencial" (Oportunidad)
Perfil: Son clientes "nuevos" para la campaña (sin contactos previos) pero con características socioeconómicas muy similares al grupo exitoso (Clúster 1): saldo saludable (1452€) y trabajos de gestión.
Comportamiento: Tienen una tasa de conversión decente (12.09%), superior al promedio general del banco.
Acción Estratégica: Captación Prioritaria. Representan la mejor oportunidad de crecimiento "en frío". Al tener capacidad de ahorro (balance) y estabilidad laboral, son el target ideal para asignar a los ejecutivos de venta telefónica.

Clúster 1: "Clientes Reactivos / Fidelizados" (La Joya)

Perfil: Este es el grupo más distintivo. Se caracteriza por tener un historial de interacciones previas con el banco, su perfil demográfico es similar al promedio (edad ~41, saldo medio-alto), y ocupan mayoritariamente cargos de gestión (management).
Comportamiento: Poseen la Tasa de Conversión más alta (22.72%).
Acción Estratégica: Prioridad Máxima. Estos clientes ya conocen el banco y han reaccionado positivamente. La estrategia debe ser de fidelización y venta cruzada (cross-selling). El costo de adquisición es bajo porque ya existe una relación.

Clúster 2: "Segmento de Baja Propensión" (Riesgo)
Perfil: Clientes sin historial previo, con el saldo promedio más bajo (1146€) y una mayor concentración en trabajos manuales o técnicos (blue-collar). Es el grupo demográficamente más joven (39 años).
Comportamiento: Tienen la Tasa de Conversión más baja (5.58%).
Acción Estratégica: Eficiencia de Costos. Contactar a este grupo por teléfono es costoso e ineficiente dado su bajo retorno. Se recomienda utilizar canales pasivos y de bajo costo (Email Marketing, SMS, Notificaciones App) y no gastar recursos de call center aquí a menos que el modelo predictivo asigne una probabilidad individual muy alta.

5. Análisis del modelo baseline

El modelo de Regresión Logística, configurado con class_weight='balanced', establece la línea base de desempeño con los siguientes hallazgos:

- ROC-AUC (0.7702): El modelo tiene una capacidad predictiva aceptable (superior al 0.5 del azar). Este es el número para vencer los modelos avanzados (SVM, RF, XGBoost).
- Recall de la Clase 1 (63%):
    - Interpretación: De todos los clientes que realmente contratarían el depósito, el modelo es capaz de detectar al 63% (997 de 1587).
    - Impacto: Es un buen número para captación. Significa que perdemos el 37% de las oportunidades, pero capturamos la mayoría.
- Precisión de la Clase 1 (27%):
    - Interpretación: De cada 100 personas que el modelo dice "Llama a este cliente", solo 27 contratarán.
    - Costo: Esto implica que el equipo de ventas hará muchas llamadas "infructuosas" (Falsos Positivos = 2720) para encontrar a los clientes reales.
- Trade-off: Al usar pesos balanceados, hemos sacrificado precisión para ganar cobertura (Recall). En marketing, esto suele ser preferible si el costo de realizar una llamada es bajo comparado con el beneficio de ganar un cliente.

El modelo es funcional y útil para filtrar la base de datos, pero tiene un costo operativo alto debido a los Falsos Positivos. Los modelos no lineales (Random Forest, SVM) deberían mejorar la precisión sin sacrificar tanto recall.

Modelo XGBoost

Se desarrolló un modelo de inteligencia artificial capaz de ordenar a los clientes por su probabilidad de compra.

- Capacidad de Detección: El modelo es capaz de identificar al 64% de los clientes que contratarán el producto antes de realizar la campaña.
- Impacto Operativo: Al utilizar este modelo, el banco puede enfocar sus recursos solo en los clientes con alta probabilidad, evitando molestar a quienes no tienen interés y reduciendo el tiempo improductivo del Call Center.

SVM Optimizado con GridSearchCV

Debido a la naturaleza no lineal del kernel utilizado en SVM, la interpretabilidad directa de coeficientes no es posible. Se aplicó el método de Importancia por Permutación. Este método evalúa la importancia de una variable mezclando aleatoriamente sus valores y midiendo la caída en la métrica ROC-AUC. Si el rendimiento del modelo empeora drásticamente al 'romper' una variable, se concluye que dicha variable es crítica para la predicción.

Random Forest con GridSearchCV

La importancia de las características en el modelo Random Forest se calculó midiendo cuánto reducen las variables la impureza del modelo (Gini) en todos los árboles del bosque. Cada vez que una variable se utiliza para dividir un nodo, se calcula la mejora en la homogeneidad de los sub-nodos resultantes. Una variable con mayor importancia indica que es fundamental para separar las clases (clientes que compran vs. los que no) de manera pura."

Gradient Boosting con GridSearchCV

Se implementó una estrategia de búsqueda exhaustiva de hiperparámetros (GridSearchCV). A diferencia de Random Forest, GBM es altamente sensible a la configuración de sus parámetros, por lo que el ajuste fino es crítico para evitar el overfitting (sobreajuste).

Al igual que en los otros modelos, variables derivadas del contacto previo (como poutcome_success o duration si se incluyó) suelen dominar. Esto indica que la historia reciente del cliente es el mejor predictor del futuro.

6 Comparación de modelos

La comparación realizada en el notebook adjuntado revela una clara superioridad de los métodos de ensamble (Random Forest y XGBoost) sobre los modelos lineales y de margen (Logística y SVM). Mientras que las curvas de RF y XGBoost muestran una alta sensibilidad en los deciles superiores, la curva del SVM (morada) se mantiene peligrosamente cerca del baseline (azul), demostrando que el alto costo computacional del SVM no se traduce en un beneficio predictivo real para este problema de negocio. Esto consolida la elección de XGBoost como el modelo final por su equilibrio óptimo entre precisión (AUC ~0.80) y eficiencia operativa.

Modelo	AUC	Recall (Clase 1)
Regresión Logística (Baseline)	0.7702	0.6282
Random Forest	0.7976	0.5558
XGBoost	0.7972	0.6434
SVM	0.7776	0.5873

7 Conclusiones

Tras analizar el comportamiento de los clientes del Banco de Portugal mediante técnicas de Machine Learning, presentamos los siguientes hallazgos y estrategias para optimizar la venta de depósitos a plazo:

A. Diagnóstico de la Cartera: Identificamos que la estrategia actual de llamadas aleatorias es ineficiente. A través de la segmentación (Clustering), detectamos tres perfiles claros.

1. **Clientes Fidelizados (Clúster 1):** Alta probabilidad de compra (>22%). Deben ser la prioridad número uno.
2. **Prospectos Potenciales (Clúster 0):** Clientes nuevos con perfil financiero sólido. Oportunidad de captación.
3. **Grupo de Riesgo (Clúster 2):** Baja probabilidad (<5%). Recomendamos no asignar ejecutivos telefónicos a este grupo para reducir costos.

B. Herramienta Predictiva (Modelo XGBoost): Se ha desarrollado un modelo de inteligencia artificial capaz de ordenar a los clientes por su probabilidad de compra.

- **Capacidad de Detección:** El modelo es capaz de identificar al **64% de los clientes** que contratarán el producto antes de realizar la campaña.
- **Impacto Operativo:** Al utilizar este modelo, el banco puede enfocar sus recursos solo en los clientes con alta probabilidad, evitando molestar a quienes no tienen interés y reduciendo el tiempo improductivo del Call Center.

C. Recomendación Estratégica: Implementar el modelo XGBoost **en producción para:**

1. **Filtrar las bases de datos** semanalmente antes de asignarlas a los ejecutivos.
2. Utilizar el **"Clúster ID"** para personalizar el guion de venta (ej. ofrecer beneficios de fidelidad al Clúster 1 vs. beneficios de entrada al Clúster 0).
3. Descartar el uso de modelos lineales simples o SVM debido a su menor rendimiento y altos tiempos de cómputo, garantizando así una operación ágil y escalable.