Informe Técnico: Limpieza y Optimización de Datos De Clientes contactados en Campaña de Venta de Certificado de Depósito a Término

Autor: Alfredo Ruiz S.

Informe de Análisis de Datos

Problema:

Una entidad bancaria contrata a una empresa de marketing encargada de contactar telefónicamente a posibles clientes para determinar si están interesados o no en adquirir un certificado de depósito a término con el banco.

Preguntas a resolver

- ¿Cuál es la rentabilidad neta de la campaña actual? (Visualización de Profit total y ROI global).
- ¿Cuánto dinero estamos "quemando" en llamadas ineficientes? (Costo acumulado de llamadas fallidas vs. exitosas).
- ¿Cuál es el perfil del cliente de alto valor? (Segmentación por balance, tipo de trabajo, estudios y si tiene o no créditos).
- ¿En qué punto deja de ser rentable insistir con un cliente? (Análisis del ROI por número de intentos/campaña).
- ¿Cómo afecta la estacionalidad a nuestra tasa de conversión? (Identificación de meses de "cosecha" vs. meses de saturación).
- ¿Cuál es el impacto financiero de aplicar la "Estrategia Híbrida"? (Comparativa de ahorro en OPEX vs. pérdida de alcance de clientes).

Resultados para las partes Interesadas

Recomendación:

Aplicar todos los filtros de manera simultánea reduciría la base de datos de 45,189 a solo 8,861 registros. Si bien esto es "hiper-eficiente", pone en riesgo el Brand Awareness y la captación de futuros clientes.

Optar por un modelo híbrido específicamente mi recomendación 2 que sería excluir saldos ≤ 0 (salvando a entrepreneur (priorizando este) y self-employed) y limitar a 3 intentos de llamada. Esta estrategia logra una reducción de costos del 29.91% manteniendo una base amplia de 30,336 clientes potenciales. Esto garantiza la salud financiera de la campaña hoy, sin sacrificar la cuota de mercado ni la visibilidad del banco a futuro.

VALIDACIÓN

Introducción

Este es un informe técnico detallado del proceso de limpieza y preparación de datos realizado en el notebook Venta_de_derecho_de_deposito.ipynb. Como analista de datos, he desglosado el flujo de trabajo en etapas lógicas, desde la ingesta hasta la generación de métricas de negocio.

Inspección Inicial

Tamaño Original: 45.215 filas y 17 columnas

Título de la columna	Descripción
age	Edad(numérica)
job	Tipo de trabajo (categórica: “admin.”, “unknown”, “unemployed”, “management”, “housemaid”, “entrepreneur”, “student”, “blue-collar”,”self-employed”, “retired”, “technician”, “services”)
marital	estado civil (categórica: “married”, “divorced”, “single”)
education	nivel educativo (categórica: “unknown”, “secondary”, “primary”, “tertiary”)
default	Si dejó de pagar sus obligaciones (categórica: “yes”, “no”)
balance	saldo promedio anual en euros (numérica)
housing	¿tiene o no crédito hipotecario? (categórica: “yes”, “no”)
loan	¿tiene créditos de consumo? (categórica: “yes”, “no”)
contact	medio a través del cual fue contactado (categórica: “unknown”, “telephone”, “cellular”)
day	último día del mes en el que fue contactada (numérica)
month	último mes en el que fue contactada (categórica: “jan”, “feb”, “mar”, …, “nov”, “dec”)
duration	duración (en segundos) del último contacto (numérica)
campaign	número total de veces que fue contactada durante la campaña (numérica)
pdays	número de días transcurridos después de haber sido contactado antes de la campaña actual (numérica. -1 indica que no fue contactado previamente)
previous	número de veces que ha sido contactada antes de esta campaña (numérica)
poutcome	resultado de la campaña de marketing anterior (categórica: “unknown”, “other”, “failure”, “success”)

y	categoría ¿el cliente se suscribió a un depósito a término? (categórica: “yes”, “no”)

Problemas identificados:

- Nulos: 8 aproximado en algunas columnas (algunas son menos) siendo ±0.0177%
- Duplicados: 4 filas duplicadas
- Outliers: múltiples valores fuera de rango como en las columnas age valores entre 300 a 500 años, balance > 500.000 (este ultimo solo afecta en la media) y valores en duration negativos
- Formatos inconsistentes como admin., string con letras mayúsculas o minúsculas que no coincidían entre ellas como MANAGEMENT vs management
- Filas totales después de la limpieza: 45189 y 21 columnas se agregaron

Columnas agregadas

Título de la columna	Descripción
Is_success	Para binarización de la columna Y, (categórica: 1, 0)
Call_cost	Costo de la llamada por cliente
Profit	Rentabilidad o pérdida de cada llamada individual
ROI	Retorno de la Inversión por llamada

Definición de las nuevas columnas:

- Is_success es una variable binaria (0 o 1) que se creó para representar de forma numérica si la campaña de marketing tuvo éxito o no con un cliente en particular para simplificar el codigo. Su definición es simple:
  Si el valor en la columna Y (que probablemente significa 'yes' o 'no' para el resultado de la campaña) es 'yes', entonces Is_success se establece en 1.
  Si el valor en la columna Y es 'no', entonces Is_success se establece en 0.
  En resumen, 1 significa que la llamada fue exitosa en lograr una suscripción, y 0 significa que no lo fue.

- Call_cost (Costo de la Llamada) representa el costo operativo estimado de cada llamada de campaña. Se calcula basándose en la duración de la llamada y el costo por minuto establecido:
  Call_cost = (Duration / 60) * call_cost
  Donde:
  Duration: Es la duración de la llamada en segundos.
  60: Es el factor para convertir los segundos a minutos.
  call_cost: Es el costo operativo por minuto de la llamada (establecido en $4.5).

- Profit: La columna Profit (Ganancia) calcula la rentabilidad o pérdida de cada llamada individual. Se determina de la siguiente manera:

Si la llamada fue exitosa (Is_success es 1), se le asigna la ganancia establecida por CDT (margin_cdt, que se estableció en $50). A esa ganancia (o cero si no fue exitosa), se le resta el costo de la llamada (Call_cost). Entonces, la fórmula es: Profit = (Is_success * margin_cdt) - Call_cost.

- ROI: La columna ROI (Return on Investment, o Retorno de la Inversión) indica la eficiencia de cada llamada en términos de rentabilidad. Se calcula usando la siguiente fórmula:
  ROI = (Profit / Call_cost) * 100
  En otras palabras, el ROI mide cuánta ganancia se obtuvo por cada unidad de dinero gastada en la llamada, expresado como un porcentaje.
  Un ROI positivo significa que la llamada generó más dinero del que costó (es rentable).
  Un ROI negativo (como -100% cuando Is_success es 0) significa que la llamada fue una pérdida, y el costo de la llamada no fue compensado por ninguna ganancia.

Pasos de Limpieza

Fase de Configuración e Ingesta

El proceso comienza con la importación de las siguientes librerías de Python: pandas para la manipulación, matplotlib y seaborn para la visualización, y os para la gestión de rutas.

- Acción: Se monta Google Drive para acceder al dataset CSV y se realiza una carga inicial en un DataFrame (df_full) desde Google Drive.
- Inspección Inicial: Se ejecutan comandos de diagnóstico (.shape, .info(), .head()) para entender la estructura de 45,211 registros y 17 variables.

Integridad Estructural y Valores Faltantes

- Manejo de Nulos: Se identificó la proporción de valores nulos mediante df_full.isna().mean(). Al ser un dataset de marketing bancario donde la precisión es clave, se optó por la eliminación de registros con valores faltantes mediante dropna(inplace=True).
- Eliminación de Duplicados: Se utilizó drop_duplicates() para garantizar que cada interacción cliente-llamada sea única, evitando sesgos en el cálculo de la tasa de conversión.

Tratamiento de Valores Atípicos (Outliers)

Esta fase es crítica para evitar que datos erróneos distorsionen los promedios de éxito.

- Visualización: Se emplearon Boxplots para identificar valores extremos en variables numéricas como age, balance y duration.
- Filtros de Lógica de Negocio:
    - Edad: Se restringió el rango a (18, 100) años, eliminando registros imposibles (como edades de 530 años detectadas en la exploración).
    - Duración: Se eliminaron llamadas con duración ≤ 0 segundos, ya que no representan una interacción real.
    - Contactos Previos: Se limitó la variable previous a ≤ 100 para eliminar ruidos estadísticos de campañas masivas ineficientes.

Normalización Categórica y Corrección Tipográfica

El código aborda el problema de la "suciedad" en los datos de texto (Case sensitivity y sinonimia):

- Estandarización de Formato: Se aplicó .str.lower() a todas las columnas categóricas para unificar valores como "Management", "management" y "MANAGEMENT".
- Mapeo de Sinónimos: Se realizaron reemplazos específicos para consolidar categorías:
    - admin. → administrative
    - div. → divorced
    - sec. → secondary
    - unk. → unknown
    - phone/mobile → telephone/cellular
- Limpieza de Cabeceras: Se aplicó strip() y title() a los nombres de las columnas para asegurar una presentación profesional y facilitar el llamado de variables en el código posterior.

Ingeniería de Características (Feature Engineering)

Para transformar un dataset de limpieza en uno de análisis estratégico, se crearon nuevas métricas financieras:

- Variable Objetivo: Is_success (binarización de la columna y).
- Modelado de Costos: Se integraron variables externas de negocio (Costo por minuto de llamada = $4.5, Margen de ganancia por CDT = $50).
- Cálculo de Rentabilidad: Se generaron las columnas Call_cost, Profit y ROI, permitiendo evaluar no solo si el cliente compró, sino cuánto costó conseguir esa venta.

Análisis Estratégico y Conclusiones

El proceso finaliza con una segmentación avanzada para derivar las recomendaciones:

- Análisis de Umbral: Comparación de medias de duración entre éxitos y fracasos.
- Segmentación por Perfil: Cruce de variables Job y Balance para identificar nichos rentables (como los emprendedores con saldo negativo que, contra todo pronóstico, tienen alta conversión).
- Optimización de Campaña: Se simularon escenarios de filtrado (ej. solo balances positivos, máximo 3 contactos) para calcular la reducción del gasto operativo (OPEX).

Código

Ver en github.com:
https://github.com/Santacaterina/data_analysis_project/tree/main/Venta_derecho_deposito

Conclusiones

El dataset ahora renombrado como dataset_banco_cleaned.csv es ahora confiable para análisis en un dashboard interactivo para su presentación para así tomar decisiones, pero se llegaron a múltiples hallazgos:

- Se identifica un umbral crítico de conversión en la duración de las llamadas. Existe una correlación positiva entre el tiempo de permanencia y el éxito de la venta: el promedio de éxito se sitúa en 537 segundos (9 min) frente a los 221 segundos (3.7 min) de las llamadas fallidas.
  Recomendación Estratégica: Implementar un protocolo de "Salida Inteligente" a partir de los 12 minutos. En lugar de un corte rígido, se debe capacitar a la fuerza de ventas en la detección de leads analíticos (aquellos que requieren más información, pero muestran interés) vs. leads pasivos. Reducir la fricción en llamadas prolongadas sin señales de cierre optimizará el ROI por hora-hombre.

- El análisis de conversión según el estado financiero revela que los clientes con balance negativo o cero tienen una Tasa de Conversión (CR) del 4.83% (excluyendo nichos específicos), situándose por debajo del benchmark bancario del 5.5%. No obstante, detectamos un comportamiento atípico en los segmentos 'entrepreneur' (11.50% de CR) y 'self-employed' (7.95% de CR), quienes muestran resiliencia incluso con balances negativos.
  Recomendación Estratégica: Aplicar un filtro de exclusión para balances ≤ 0, con excepción de los segmentos entrepreneur y self-employed. Se recomienda descartar el segmento 'student' con saldo negativo, ya que su tasa de conversión es nula (0%).

- Existe una correlación directa y positiva entre el volumen de activos en cuenta (balance) y la probabilidad de suscripción. Los clientes con mayor liquidez presentan una menor resistencia a la inmovilización de capital en productos de ahorro a plazo.

- La ausencia de pasivos hipotecarios duplica la probabilidad de contratación del CDT. Esto se debe a una mayor capacidad de ahorro discrecional y menor carga financiera mensual.
  Recomendación Estratégica: Priorizar en el modelo de lead scoring a los clientes sin hipoteca para maximizar el volumen de captación, pero manteniendo un seguimiento diferenciado para clientes con hipoteca para fomentar la fidelización y el cross-selling de largo plazo.

- Se observa una estacionalidad marcada con picos de éxito >40% en marzo, septiembre, octubre y diciembre. Por el contrario, mayo y julio muestran una ineficiencia operativa: alto volumen de llamadas con baja tasa de retorno.
  Diagnóstico: Estos meses podrían estar sufriendo de fatiga de contacto o saturación de campañas concurrentes.
  Recomendación Estratégica: Redistribuir el presupuesto de marketing hacia los meses de alta conversión y realizar un análisis de causa raíz sobre el bajo rendimiento de mayo para ajustar el script o la oferta comercial.

- El análisis de frecuencia indica que la probabilidad de éxito marginal se desploma después del tercer intento. Realizar más de 3 llamadas al mismo cliente resulta en un ROI negativo, ya que el costo operativo supera la probabilidad estadística de cierre.
  Recomendación: Establecer un límite de contacto (Hard Cap) de 3 intentos por campaña.

pero se llegó a múltiples conclusiones y recomendaciones aplicando una serie de estrategias a partir de los hallazgos donde se crea otro dataset llamado df_opt (como base ya se derivaron otros dataset a partir de este) donde se descontó los clientes con valores en 0 o negativo en su balance, no más de 3 llamadas y un que no hayan dejado de pagar sus compromisos llegando a estas conclusiones:

1- Solo llamando a cliente con balance >0, y no más de 3 llamadas:
   La exclusión técnica de cuentas con saldo cero o negativo representa una reducción inmediata del 31.23% en el Gasto Operativo (OPEX), permitiendo a la agencia centrar sus recursos en segmentos con mayor capacidad de inversión.

2- Esta estrategia no varía mucho de la anterior a diferencia que se agrega a los entrepreneur y self-employed debido que a pesar de tener un balance ≤0 se tiene una aceptación de los certificados de depósito dando una reducción de los gastos operación en un 29.91% pero aumenta el número de clientes a llamar

3- Si se restringe la campaña únicamente a los perfiles laborales con histórico de ROI positivo (excluyendo blue-collar, services, etc.), se logra una máxima eficiencia en costos, aunque esto implica una reducción del alcance total del embudo de ventas.

4- Limitar la operación exclusivamente a los meses de mayor aceptación (Feb, Mar, Abr, Jun, Ago, Sep, Oct, Dic) permitiría una reducción drástica de costos operativos del 90.46%, convirtiendo la campaña en una operación quirúrgica de alta rentabilidad.

5- Aplicando todos los filtros o estrategias se tiene una reducción de los gastos operativos en un 92.74% que eso es un gran valor, pero solo se llega a pocos clientes

Entonces como ya se dijo al aplicar todos los filtros de manera simultánea reduciría la base de datos de 45,189 a solo 8,861 registros. Si bien esto es "hiper-eficiente", pone en riesgo el Brand Awareness y la captación de futuros clientes.

Optar por un modelo híbrido específicamente mi recomendación 2 que sería excluir saldos ≤ 0 (salvando a entrepreneur (priorizando este) y self-employed) y limitar a 3 intentos de llamada. Esta estrategia logra una reducción de costos del 29.91% manteniendo una base amplia de 30,336 clientes potenciales. Esto garantiza la salud financiera de la campaña hoy, sin sacrificar la cuota de mercado ni la visibilidad del banco a futuro.