**Informe EDA — Campaña comercial del banco**

**Autor:** Claudio Baldini

**1. Introducción**

El presente trabajo tiene como objetivo realizar un **Análisis Exploratorio de Datos (EDA)** sobre un dataset real de campañas de marketing telefónico de un banco portugués. La variable objetivo es **suscripcion** (antes llamada y), que indica si el cliente aceptó o no un depósito a plazo.

El propósito del EDA no es construir un modelo predictivo, sino **comprender la estructura de los datos, su calidad, las transformaciones necesarias y los factores más relevantes asociados al éxito de la campaña**.

**2. Metodología de trabajo**

El proyecto se desarrolló en varias etapas claramente diferenciadas:

**2.1. Inspección inicial**

* Carga de datasets:
  + **bank-additional.csv** (información de campañas).
  + **customer-details.xlsx** (información de clientes, en 3 hojas).
* Revisión de shape, tipos de datos, nulos y primeras filas.
* Detección de valores especiales como 'unknown' o espacios en blanco.

**2.2. Limpieza y transformación del dataset del banco**

* **Nombres de columnas** normalizados a snake\_case.
* **Tratamiento de pdays=999**:
  + Este valor significa “no contactado previamente”.
  + Creé una bandera (pdays\_no\_prev\_contact) y dejé pdays como numérica con NaN en esos casos.
* **Conversión de tipos**: se revisaron columnas que eran numéricas pero estaban como texto (con comas en lugar de puntos decimales).
* **Duplicados**: eliminados si había coincidencia exacta de todas las columnas.
* **Outliers**: no se eliminaron, pero se marcaron con banderas:
  + Por reglas simples (age<18, age>95, duration>2000, campaign>20).
  + Por IQR para age, duration, campaign y previous.
  + Creé banderas en dichas columnas para marcar cuando hayan valores que se alejen mucho del rango de los promedios
* **Educación**:
  + Se limpiaron nombres (basic.9y → basic\_9y).
  + Se creó la columna education\_level, agrupando en categorías: *basic*, *high\_school*, *university*, *professional*, *illiterate*.

**2.3. Limpieza y unificación del dataset de clientes**

* Por cada hoja del Excel:
  + Se normalizaron nombres de columnas.
  + dt\_customer se convirtió a tipo fecha.
  + Se generó id\_str como identificador limpio.
  + Se añadió cohort para saber de qué hoja provenía.
* Se unieron las 3 hojas con concat → **clients\_all**.
* Se revisaron nulos, duplicados por id\_str y estadísticas básicas.

**2.4. Merge final**

* Se hizo un **LEFT join** de campañas con clientes (clave: id\_str).
* Se obtuvo el dataset final **df\_master**, base para el análisis.

**3. Decisiones sobre nulos**

En un EDA descriptivo decidí **no imputar nulos**:

* pdays → queda con ~96% NaN tras eliminar el código 999. Tiene sentido: la mayoría nunca había sido contactada antes.
* Variables macro (euribor3m, emp\_var\_rate, nr\_employed) y default presentan nulos moderados.
* Justificación: **mantener los nulos permite describir fielmente el dataset**, sin inventar valores ni introducir sesgos.

**4. Análisis descriptivo y descubrimientos**

**4.1. Distribución global de la respuesta**

* La mayoría de registros son **“no”**.
* Hay pocos “yes” → el dataset está **desbalanceado**.

**4.2. Tendencia temporal (% de yes por año)**

* 2012 → ~4.6% éxito.
* 2013 → ~7.7%.
* 2014 → ~23%.  
   El éxito **aumenta con el tiempo**, especialmente en 2014.

**4.3. Variables numéricas**

* **Age**: la mayoría de clientes entre 30–40 años. Pocos >70.
* **Duration**: clave → los *yes* tienen duraciones mucho más largas.
* **Campaign**: *yes* suelen necesitar menos contactos; *no* acumulan muchos intentos.
* **Income**: no se observan diferencias claras entre *yes* y *no*.

**4.4. Variables categóricas**

* **Marital**: los *single* tienen mayor % de éxito que casados/divorciados.
* **Job**: destacan *retired* y *student* con mayor % de yes; *blue-collar* y *services* con menos.
* **Education**: diferencias leves; *university* y *illiterate* algo más positivos.
* **Housing y Loan**: sin impacto claro.

**4.5. Grupos de edad (bins)**

* Más propensos: **jóvenes (18–24)** y **mayores (>60)**.
* Menos propensos: 30–50 años.

**4.6. Correlaciones**

* Variables macro (euribor3m, emp\_var\_rate, nr\_employed) altamente correlacionadas entre sí → redundancia.
* duration no correlaciona con otras “reales” → aporta información única.
* age, income, housing, loan → correlaciones bajas.

**5. Interpretación de los gráficos (resumen)**

* **Boxplot Duration vs suscripción**: los *yes* tienen medianas y dispersión mucho mayores.
* **Violinplot Duration**: muestra que la densidad de *yes* se concentra en duraciones largas; los *no* en cortas.
* **Campaign vs suscripción**: muchos intentos → más “no”.
* **Barras Education\_level**: leve ventaja en *university* y *illiterate*.
* **Barras Job**: retirados y estudiantes responden mejor.
* **Barras Marital**: solteros responden mejor.
* **Barras Housing/Loan**: sin diferencias.
* **Age bins**: extremos de edad más favorables.
* **Heatmap**: confirma redundancias macroeconómicas y la independencia de duration.

**6. Conclusiones finales**

* **Más probabilidad de “sí”**:
  + Llamadas largas.
  + Pocos intentos de campaña.
  + Jóvenes (18–24) y mayores (60+).
  + Solteros.
  + Retirados y estudiantes.
* **Más probabilidad de “no”**:
  + Edades medias (30–50).
  + Trabajos *blue-collar* y *services*.
  + Campañas con demasiados intentos.
* **Variables macro**: aportan contexto económico, pero son redundantes entre sí.
* **Insight clave**: el **proceso de la llamada pesa más que el perfil del cliente**.

**7. Recomendaciones**

* **Optimizar la calidad de las llamadas**:
  + Priorizar llamadas efectivas, no cortas.
  + Evitar insistencia excesiva (contactar demasiadas veces al mismo cliente reduce la tasa de éxito).
* **Segmentación de campañas**:
  + Foco en jóvenes, mayores, solteros, estudiantes y retirados.
  + Menos recursos en franjas medias y perfiles resistentes.
* **Variables macro**: basta con utilizar una como proxy (ej. euribor3m).
* **Datos futuros**: registrar más información del proceso de llamada (hora, día, agente, script) para entender qué prolonga llamadas de calidad.

Con este EDA se logró:

* Transformar y limpiar los datos.
* Comprender la estructura del dataset.
* Identificar variables clave.
* Extraer insights útiles para negocio.