1. Introducción

Las campañas de marketing telefónico son un canal habitual para la comercialización de productos bancarios. En este proyecto se analiza el comportamiento de clientes contactados y los factores asociados a la suscripción de un depósito a plazo, con el fin de comprender patrones de respuesta y establecer una base sólida para tomar decisiones comerciales y para el desarrollo de modelos predictivos.

Las campañas de marketing telefónico siguen siendo un canal relevante para la comercialización de productos bancarios, especialmente cuando se requiere segmentar y priorizar contactos de forma eficiente. En este proyecto se analiza el comportamiento de los clientes contactados por una entidad bancaria y los factores asociados a la suscripción de un depósito a plazo, con el propósito de comprender los patrones de respuesta y detectar perfiles con mayor probabilidad de conversión.

El estudio se apoya en información demográfica, socioeconómica y macroeconómica, junto con variables relacionadas con el historial de campañas y la interacción con el cliente. A partir de estas fuentes, se desarrolla un análisis exploratorio estructurado que incluye una revisión inicial de la calidad de los datos, un proceso de depuración y estandarización, la integración de los datasets y la creación de variables derivadas con enfoque temporal.

Finalmente, se aplican análisis univariantes y bivariantes para identificar relaciones relevantes con la variable objetivo y extraer conclusiones útiles para la mejora de la segmentación, la optimización de recursos comerciales y la base analítica de futuras fases de modelización.

2. Objetivos del proyecto

Objetivo general

El objetivo principal del proyecto es desarrollar un análisis exploratorio de datos que permita comprender en profundidad la estructura, la distribución y las relaciones existentes entre las variables que describen a los clientes y al contexto de las campañas de marketing telefónico. A través de este estudio se pretende identificar patrones relevantes y factores asociados a la suscripción del depósito a plazo, diferenciando comportamientos entre clientes que contratan y quienes no lo hacen.

De manera adicional, el proyecto contempla la construcción de un dataset unificado y depurado a partir de las fuentes disponibles, garantizando su coherencia y calidad para el análisis. Esta base integrada permite extraer conclusiones operativas útiles para la planificación comercial, al mismo tiempo, dejar preparado un punto de partida sólido para futuras fases de modelización, enfocadas en identificar perfiles prioritarios, optimizar el orden y enfoque de los contactos y aumentar el rendimiento de las próximas campañas.

Objetivos específicos

Análisis preliminar

En esta primera fase se realiza una revisión inicial de las bases de datos con el fin de entender su estructura y contenido antes de aplicar transformaciones. Esta fase incluye la identificación de tipos de variables y formatos, el reconocimiento de posibles inconsistencias, anomalías o errores de registro, así como la detección de valores ausentes y duplicados. El objetivo es anticipar los principales problemas de calidad de los datos y definir las acciones necesarias para garantizar que el análisis posterior se realice sobre información fiable.

Limpieza, transformación e integración de datos

En esta fase se procede a depurar y estandarizar los conjuntos de datos mediante la corrección de inconsistencias, el tratamiento de valores faltantes, la eliminación de duplicados y la supresión de variables que no aporten valor analítico. Además, se aplican transformaciones orientadas a mejorar la interpretación de los datos, incluyendo recodificaciones, ajustes de formato y la creación de variables derivadas que faciliten el análisis y la comparación entre grupos.
En esta fase también se procede a la unificación de las fuentes *bank-additional.csv* y *customer-details.xlsx*, asegurando la coherencia entre claves y la consistencia en el dataset resultante. Como resultado, se construye **df_final**, base de datos integrada utilizada como referencia para el análisis exploratorio.

Análisis exploratorio de datos

En esta última fase se Analiza de forma detallada las variables y sus relaciones mediante técnicas univariantes y bivariantes, evaluando distribuciones, desbalances y posibles valores atípicos, así como su asociación con la variable objetivo. Esta fase busca identificar patrones, tendencias y factores potencialmente influyentes en la suscripción del depósito, con especial atención a las diferencias entre segmentos de clientes y al papel de variables operativas y temporales en el rendimiento de la campaña.

3. Metodología

Enfoque general del estudio

El proyecto se ha desarrollado siguiendo un flujo analítico estructurado, desde la comprensión inicial de las fuentes hasta la obtención de hallazgos interpretables sobre la variable objetivo **y**.

El análisis se ha planteado para extraer conclusiones útiles en la optimización de las campañas, al mismo tiempo, que se deja el conjunto de datos estructurado y consistente para su uso en posteriores modelos predictivos.

Criterios de limpieza, transformación e integración

Para garantizar la calidad del análisis, se aplicaron criterios consistentes en todas las transformaciones:

- **Normalización y coherencia de tipos de dato**: conversión de variables a su tipo correcto según contenido, numéricas, categóricas y temporales, asegurando consistencia en fechas y codificaciones.
- **Tratamiento de valores faltantes y duplicados**: identificación y corrección según el caso, imputación cuando era pertinente, verificación de coherencia y eliminación si procedía, además de controles posteriores para confirmar ausencia de nulos y duplicidades relevantes.
- **Estandarización de categorías**: unificación del formato de variables categóricas y sus niveles para evitar inconsistencias en cálculos y visualizaciones.
- **Eliminación de variables no informativas**: retirada de atributos que actúan como identificadores o no aportan valor explicativo al objetivo del análisis.
- **Integración de fuentes**: consolidación de bank-additional.csv y customer-details.xlsx mediante clave común, generando el dataset unificado df_final. Tras la unión se realizaron validaciones de integridad, registros esperados, claves, nulos, duplicados y coherencia general.
- **Persistencia del dataset**: almacenamiento en formatos adecuados para el flujo de trabajo, incluyendo una versión final con variables derivadas del EDA para mantener consistencia en los análisis.

Enfoque de análisis exploratorio y comparación respecto a y

El análisis exploratorio de datos se planteó con doble objetivo: describir el comportamiento global del dataset y evaluar diferencias entre clases de la variable objetivo y.

Análisis univariante

- **Numéricas**: revisión de estadísticos descriptivos, forma de la distribución y detección de outliers mediante histogramas y boxplots. Se evaluó dispersión, sesgo y valores extremos.
- **Categóricas**: análisis de frecuencias y porcentajes para identificar categorías predominantes y evaluar posibles desequilibrios en la distribución.

Análisis bivariante orientado a y

- **Numéricas vs y**: comparación de distribuciones por clase mediante boxplots y contraste de tendencia central usando medianas por grupo. Se identificaron variables con mayor separación entre clases.

- **Categóricas vs y**: uso de tablas de contingencia normalizadas por categoría para estimar tasas de respuesta, complementadas con gráficos de barras apiladas para facilitar la comparación visual entre grupos.

Correlaciones

- Construcción de matriz de correlación y mapa de calor para detectar relaciones lineales y posibles dependencias entre variables numéricas, prestando especial atención a conjuntos de variables con comportamiento conjunto.

Análisis temporal

- Creación de variables derivadas a partir de fechas para estudiar la evolución por periodos y la relación con y:
    - Contactos por año y mes.
    - Altas de clientes por año y mes.
    - Análisis de antigüedad del cliente y su relación con la probabilidad de contratación, agrupaciones mediante intervalos.

Aspectos relevantes en el informe

- **Consideración del desbalance de clases**: desde el inicio se tuvo en cuenta que, la variable y, presenta una distribución desbalanceada, lo que afecta la interpretación de tasas y la futura evaluación de modelos.
- **Trazabilidad del proceso**: el trabajo se organizó por fases y notebooks, documentando decisiones y verificaciones tras cada bloque de transformación para asegurar consistencia del dataset final.
- **Enfoque interpretativo**: además de reportar patrones estadísticos, se priorizó explicar su significado en términos de campaña segmentos, conectando hallazgos con posibles acciones.

4. Datos utilizados

Fuentes de información

El análisis se basa en la integración de dos fuentes principales:

- **bank-additional**: dataset de campañas de marketing telefónico con información del contacto, características del cliente y variables de contexto macroeconómico, además de la variable objetivo asociada a la contratación del depósito.
- **customer-details**: dataset complementario con variables socioeconómicas y de comportamiento del cliente, como ingresos, composición del hogar, fecha de alta y actividad web.

La combinación de ambas fuentes permite disponer de una visión más completa del cliente y del entorno en el momento de la campaña.

Tras los procesos de revisión, limpieza e integración, se construye un dataset unificado.

- **df_final**: conjunto de datos integrado a partir de ambas fuentes, depurado y estandarizado para su análisis.
- **df_final_eda**: versión extendida del dataset final que incorpora variables derivadas creadas durante el análisis exploratorio, especialmente orientadas al análisis temporal.

El dataset resultante incluye variables de cuatro bloques principales:

- **Demográficas y socioeconómicas**: edad, nivel educativo, ocupación, ingresos y situación familiar.
- **Operativas de campaña**: canal de contacto, duración de la llamada, número de contactos, resultado previo de campañas.
- **Macroeconómicas**: indicadores como euribor3m, emp.var.rate, cons.price.idx y cons.conf.idx.
- **Temporales:** derivadas del año y mes del contacto, año y mes de alta, y métricas de antigüedad del cliente generadas para analizar tendencias en el tiempo.

Variable objetivo y

La variable objetivo del estudio es **y**, que indica si el cliente contrata el depósito a plazo tras la campaña.

En el conjunto de datos se observa un desbalance acusado:

- **no**: 34.911 registros, aproximadamente **88,9 %**
- **yes**: 4.361 registros, aproximadamente **11,1 %**

Este reparto implica que la contratación es un evento minoritario y condiciona tanto la interpretación de resultados como cualquier análisis posterior orientado a predicción.

5. Preparación de los datos

Revisión preliminar

En una primera etapa se realizó una revisión descriptiva de **bank-additional** y **customer-details** para comprender la estructura del dato y su calidad. Esta revisión incluyó la identificación de:

- Tipos de variables y formatos.
- Posibles inconsistencias entre contenido y tipo de dato
- Presencia de valores nulos y duplicados.
- Variables con baja utilidad analítica por su naturaleza técnica.

El objetivo fue establecer qué ajustes eran necesarios antes de integrar ambas fuentes y asegurar que el análisis posterior se realizara sobre un conjunto consistente.

Limpieza y transformaciones aplicadas

A partir de los hallazgos de la revisión preliminar se aplicaron tareas de depuración y estandarización orientadas a dejar el dato preparado para el análisis:

- **Ajuste de tipos de dato**: conversión de variables al formato adecuado según su contenido, incluyendo fechas a datetime y recodificaciones coherentes en variables numéricas y categóricas.
- **Tratamiento de valores ausentes**: verificación e imputación cuando correspondía, priorizando mantener coherencia semántica de cada variable.
- **Gestión de duplicados**: comprobación de duplicidades y eliminación de registros repetidos cuando podían afectar la representatividad del dataset.
- **Estandarización de categorías**: homogeneización de etiquetas y valores para evitar categorías duplicadas por diferencias de formato.
- **Eliminación de variables no informativas**: variables puramente identificativas o técnicas se excluyeron del análisis explicativo para evitar ruido en las conclusiones.

Unión de fuentes y validaciones

Una vez depuradas las fuentes, se procedió a su integración para construir el dataset unificado:

- **Unión de bank-additional y customer-details** mediante una clave común de cliente, garantizando que la relación entre registros fuera consistente.
- **Validaciones posteriores a la unión**:
    - Comprobación de valores nulos en campos críticos.
    - Verificación de duplicados tras la integración.
    - Control de integridad de la clave utilizada en el merge.
    - Revisión del número final de registros y columnas para confirmar consistencia.

El resultado de este proceso fue **df_final**, utilizado como base principal del análisis exploratorio.

Variables derivadas creadas para el análisis temporal

Con el objetivo de estudiar los patrones y las tendencias en el tiempo, se generaron variables temporales adicionales que se incorporaron a **df_final_eda**:

- Año y mes del contacto a partir de **date**:
  - year_contact
  - month_contact
- Año y mes de alta del cliente a partir de **dt_Customer**:
  - year_customer
  - month_customer
- Tiempo transcurrido entre alta y contacto:
  - days_until_signup, diferencia temporal entre la fecha de contacto **date** y la fecha de alta del cliente **dt_customer.**
  - days_until_signup_num, misma información que days_until_signup, pero convertida a número entero, para facilitar operaciones posteriores, ya que **pd.cut()** no admite valores del tipo datetime.
- Agrupación por tramos de antigüedad:
  - time_until_signup, generada mediante pd.cut() para segmentar a los clientes en intervalos de antigüedad comparables.

Estas variables permiten estudiar la actividad en la campaña a lo largo del tiempo, identificando patrones en las altas de clientes, y así poder evaluar cómo la antigüedad se relaciona con la probabilidad de contratación.

6. Análisis exploratorio

Análisis univariante

Variables numéricas

El análisis de distribuciones y la detección de valores atípicos evidencian comportamientos distintos según la naturaleza de cada variable.

- **age**: distribución concentrada principalmente entre 30 y 50 años, con algunos valores altos poco frecuentes. Los extremos observados son plausibles y no sugieren errores evidentes.
- **duration**: distribución fuertemente sesgada a la derecha. La mayoría de llamadas son cortas y existe una cola larga con valores muy elevados, con máximos que alcanzan **4.918 segundos**. Estos outliers son coherentes con la naturaleza de la variable, pero pueden influir en medias y modelos.

- **campaign**: sesgo a la derecha con la mayor parte de clientes entre 1 y 3 contactos, y presencia de casos extremos con valores altos, llegando hasta **43** contactos.
- **emp.var.rate**: valores dentro de un rango acotado, sin outliers relevantes desde el punto de vista técnico.
- **cons.price.idx**: variable muy estable, con variabilidad baja y sin extremos destacables.
- **cons.conf.idx**: dispersión moderada, con valores en los extremos del rango que siguen siendo compatibles con el indicador.
- **euribor3m**: distribución heterogénea, con concentración en varios niveles y presencia de valores elevados que generan cola derecha, refleja cambios de régimen en el contexto económico.
- **income**: marcada asimetría a la derecha, con un núcleo central amplio y un grupo minoritario con rentas altas. Se observan valores extremos, con máximo de **180.802**, habituales en variables económicas.
- **kidhome y teenhome**: variables discretas acotadas entre 0 y 2, no admiten outliers por definición.
- **numwebvisitsmonth**: distribución moderadamente dispersa, con algunos valores altos poco frecuentes, con máximo de **32** visitas mensuales.

Variables categóricas

El análisis de frecuencias permite detectar categorías predominantes y localizar variables con distribuciones claramente desbalanceadas.

- **Job**: predominan perfiles admin., blue-collar y technician, que concentran la mayor parte de la muestra. El resto de categorías aparece con menor peso relativo.
- **education**: predominan university.degree y high.school, seguidas por niveles básicos y formación profesional. illiterate es residual.
- **housing**: distribución relativamente equilibrada entre **yes** y **no**.
- **loan**: desbalance claro, con mayoría de **no** frente a **yes**.
- **contact**: predomina cellular sobre telephone.
- **poutcome**: fuerte concentración en nonexistent, mientras que success es muy minoritaria.
- **y**: desbalance marcado de la variable objetivo, con **88,9 % no** y **11,1 % yes**, lo que condiciona la interpretación de resultados y la evaluación de modelos posteriores.
- **id**: actúa como identificador único, sin interpretación analítica en términos de distribución o patrones.

Análisis bivariante respecto a y

Variables numéricas vs y

La comparación entre clases permite identificar qué variables diferencian con mayor claridad a los clientes que contratan del depósito frente a los que no.

- **duration**: es la variable que muestra una mayor diferencia, la mediana pasa de 163 en y = no, a 453 en y = yes. Los boxplots confirman un desplazamiento claro del grupo yes hacia duraciones mayores y con mayor dispersión.
- **Entorno macroeconómico**: se observan desplazamientos consistentes en el grupo yes hacia valores más bajos:
  - **emp.var.rate**: mediana **1,1** en no, frente a **-1,8** en yes.
  - **euribor3m**: mediana **4,958** en no, frente a **1,405** en yes, siendo el contraste más marcado dentro del bloque macroeconómico.
  - **cons.price.idx**: mediana **93,918** en no, frente a **93,2** en yes.
  - **cons.conf.idx**: mediana **41,8** en no, frente a **40,8** en yes.

Las variables age, campaign, income, kidhome, teenhome y numwebvisitsmonth presentan medianas muy similares entre clases y boxplots altamente solapados, por lo que su poder discriminativo es limitado a este nivel descriptivo.

Variables categóricas vs y

Las tablas de contingencia y los gráficos de barras apiladas evidencian variaciones significativas en la tasa de contratación según la categoría de cada variable.

- **poutcome**: es la variable con mayor capacidad discriminativa.
  - **success** alcanza **64,81 % yes**.
  - **failure** registra **13,49 % yes**.
  - **nonexistent** baja a **8,80 % yes**.

  Esto indica que un historial previo positivo está fuertemente asociado a la conversión.
- **contact**: diferencias muy marcadas por canal.
  - **cellular** presenta **14,43 % yes**.
  - **telephone** se sitúa en **5,16 % yes**.

  El contacto por móvil se asocia a una mayor tasa de respuesta.
- **job**: heterogeneidad clara por ocupación.

    - Tasas altas: **student 29,58 %** y **retired 24,64 %**.
    - Tasas bajas: **blue-collar 6,98 %**, **services 7,93 %** y **entrepreneur 8,05 %**.
- **education**: patrón general favorable a niveles formativos más altos.
    - **university.degree 13,64 %** y **professional.course 11,33 %** se sitúan en niveles superiores a la tasa media, seguidos de **high.school** con **10,91%**
    - **illiterate** debe interpretarse con cautela porque, aunque presenta un **22,22%** de **yes**, solo presenta 18 valores en conjunto de datos
- **marital**: diferencias moderadas, con **single 13,57 % yes** por encima de **married 10,10 %** y **divorced 10,30 %**.
- **housing** y **loan**: diferencias muy pequeñas, sin evidencia clara de un efecto fuerte en este nivel.

Correlaciones

Matriz de correlación y mapa de calor

La matriz y el mapa de calor reflejan, en general, **correlaciones bajas**, lo que sugiere ausencia de multicolinealidad fuerte en la mayor parte del conjunto.

- La relación más destacada es **emp.var.rate con cons.price.idx**, con una correlación positiva alta de **0,78**.
- Se observa una correlación negativa moderada entre **emp.var.rate y cons.conf.idx** de **-0,21**.
- **duration** presenta correlaciones prácticamente nulas con el resto, coherente con su naturaleza operativa ligada a la llamada.
- Variables como **income**, **kidhome**, **teenhome** y **numwebvisitsmonth** se comportan de forma muy independiente respecto al resto, ya que sus coeficientes de correlación con las demás variables se mantienen muy próximos a cero.

Análisis temporal

Contactos

- **Contactos por año**: el volumen de contactos entre 2015 y 2019 es muy estable, sin cambios operativos relevantes en intensidad anual.
- **Contactos por mes**: la distribución mensual es prácticamente uniforme, sin estacionalidad marcada en el volumen de actividad.

Tasa de contratación

- **Tasa anual de contratación**: variaciones moderadas, alcanzando su máximo en **2016** alrededor del **11,7 %**, un descenso en **2017** por debajo del **10,8 %**, y mostrando una recuperación gradual en **2018** y **2019** hasta **11,18 %** y **11,19 %** respectivamente.

- **Tasa mensual de contratación:** se mantiene en un rango relativamente estrecho, con un mínimo en **septiembre** de **10,21 %** y un máximo en **octubre** de **12,25 %**. Fuera de esos extremos, las tasas mensuales se mueven con variaciones moderadas, sin un patrón estacional claramente marcado.

Altas de clientes

- **Altas por año:** el mayor volumen de incorporaciones se concentra en **2012** con **18.145 clientes**, seguido de una caída en **2013** con **8.330**, y una recuperación parcial en **2014** con **12.797**. Este comportamiento evidencia que el crecimiento de la base de clientes no fue uniforme en el periodo analizado.
- **Altas por mes acumuladas**: patrón homogéneo, con mínimo en **febrero 3.044** y máximos en **mayo 3.402** y **agosto 3.452**, sin estacionalidad intensa.

Antigüedad y probabilidad de contratación

Los resultados muestran un patrón consistente, los clientes más recientes responden mejor.

- En la segmentación por antigüedad, los grupos con menos de un año desde el alta se sitúan alrededor del **20 %** de respuesta.
- El grupo de más de 2 años desciende hasta aproximadamente el **10 %**.

Esto sugiere que la receptividad a la campaña es mayor en etapas tempranas de la relación con la entidad y disminuye conforme aumenta la antigüedad.

7. Aplicaciones prácticas para optimizar la campaña

Los hallazgos del análisis ofrecen conclusiones aplicables a la planificación y ejecución de campañas telefónicas para la suscripción del depósito, especialmente en lo relativo a la segmentación de clientes y la priorización de contactos.

Segmentos con mayor probabilidad de conversión

Se identifican perfiles con mayor propensión a contratar que pueden emplearse para mejorar la segmentación y la selección de audiencias. En particular, la tasa de conversión varía de forma relevante según características del cliente, destacando categorías con rendimientos por encima de la media como **student** y **retired**, así como segmentos con desempeño inferior como **blue-collar** o **services**. Estas diferencias sugieren que la ocupación es una variable útil para focalizar esfuerzos comerciales y adaptar el mensaje.

Además, el análisis por características formativas muestra un patrón consistente en el que niveles educativos más altos tienden a asociarse con tasas de respuesta superiores, mientras que los niveles básicos presentan tasas más bajas. Esto puede orientar campañas con propuestas de valor diferenciadas y estrategias de comunicación adaptadas al perfil.

Canales y factores operativos relevantes

El canal de contacto tiene un impacto claro en la eficacia de la campaña. El contacto mediante telefonía móvil muestra tasas de conversión significativamente superiores frente a la línea fija, lo que indica que el canal móvil resulta más eficiente para maximizar la probabilidad de respuesta positiva.

Desde un punto de vista operativo, esto respalda la recomendación de priorizar el canal móvil en la planificación de llamadas y utilizar el canal fijo de forma más selectiva.

En cuanto a los factores operativos, la duración de la llamada destaca como el indicador con mayor capacidad de diferenciación entre clientes que contratan y los que no, lo que sugiere que interacciones más extensas suelen asociarse a un mayor nivel de interés o a una mejor gestión del contacto.

Este resultado puede llevarse a medidas concretas, como optimizar los guiones de llamada, reforzar la formación de los agentes o ajustar las estrategias de seguimiento para mejorar la calidad de la interacción, más que incrementar únicamente el número de contactos.

Peso del historial previo y antigüedad

La variable **poutcome** es el factor categórico más determinante, los clientes con resultado previo **success** presentan una tasa de conversión muy superior al resto, lo que indica que el historial de respuesta positiva es un criterio de priorización especialmente valioso. Incluso cuando el resultado previo es **failure**, la tasa de contratación se sitúa por encima de la categoría **nonexistent**, lo que sugiere que el simple hecho de contar con historial aporta información útil.

La variable poutcome es el factor categórico más determinante. Los clientes con resultado previo **success** presentan una tasa de conversión del **64,81 %**, muy superior al resto, lo que confirma que el historial de respuesta positiva es un criterio de priorización especialmente valioso. Incluso cuando el resultado previo es **failure**, la tasa de contratación alcanza el **13,49 %**, por encima de **nonexistent** con **8,80 %**, lo que sugiere que contar con historial previo aporta información adicional útil, aunque no haya sido exitoso.

Al mismo tiempo, la dimensión temporal confirma que la antigüedad del cliente está asociada a la probabilidad de contratación. Los clientes más recientes muestran las tasas más altas, con aproximadamente un **20 %** de respuesta positiva tanto en el tramo de menos de **6 meses** como en el de **6** a **12 meses** desde el alta. A medida que aumenta la antigüedad, la conversión desciende, situándose en torno al **14–15 %** para clientes con **1** a **2 años** y alrededor del **10 %** para los que superan los **2 años**. Este patrón sugiere una mayor receptividad en las primeras etapas de la relación con la entidad y respalda acciones específicas de activación temprana tras el alta.

En conjunto, los resultados apuntan a que la efectividad de la campaña puede incrementarse mediante una combinación de:

- Priorización de segmentos con mayor propensión de conversión.
- Preferencia por el uso del canal móvil como vía principal de contacto.
- Uso del historial previo como criterio central de selección y priorización.

- Acciones específicas para clientes recientes, con mayor tasa de respuesta.
- Mejorar la calidad de la interacción, dado que las llamadas de mayor duración se asocian con una mayor probabilidad de contratación.

8. Conclusiones

Para finalizar podemos concluir……..