Proyecto

Entregable Final

Informe de Análisis de Datos Project SmartBank

Grupo No. 1:

Anzules Fuentes Abraham Joel

Barona Zambrano Eduardo Geovanny

Baño Cordero Christell Nicole

Mera Lopez Monica Lisbeth

Vera Espinoza Angel Isaac

Vivas Segovia Victor Augusto

Python for Data Analytics

Ing. Eduardo Cruz, PhD.

Coding Bootcamp Espol

Mintel

CBMP4

14 de nov. de 25

Informe de Análisis de Datos Project SmartBank

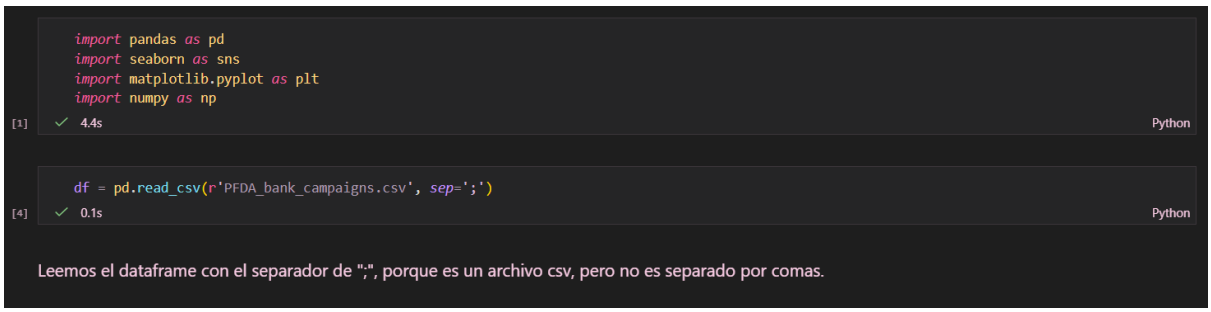
1. Introducción

El objetivo del análisis fue examinar un conjunto de datos relacionados con campañas de marketing y determinar patrones, características y comportamientos relevantes de distintos segmentos de clientes. Para ello, se hizo uso de las bibliotecas estudiadas a lo largo de este módulo: Pandas, Numpy, Matplotlib y Seaborn, siguiendo un flujo de trabajo estructurado que incluyó lectura y limpieza del dataset, exploración inicial, creación de los indicadores, segmentación de clientes, visualización de tasas de conversión e interpretación de cada uno de los resultados obtenidos.

2. Descripción Técnica de los Procedimientos Utilizados

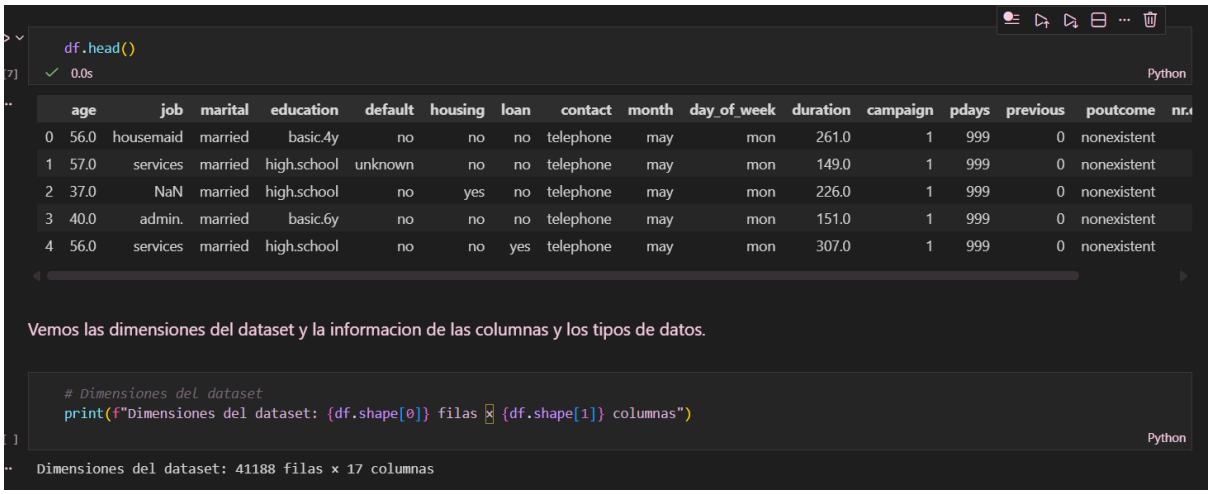
2.1 Lectura del Dataset

Se importó Pandas y se cargó el archivo *bank.csv* en un DataFrame llamado *df.*



2.2 Exploración Inicial

Con el método *head()* visualizamos las primeras filas del dataset para verificar que se cargó correctamente:



	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	nr.e
0	56.0	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261.0	1	999	0	nonexistent	
1	57.0	services	married	high.school	unknown	no	no	telephone	may	mon	149.0	1	999	0	nonexistent	
2	37.0	NaN	married	high.school	no	yes	no	telephone	may	mon	226.0	1	999	0	nonexistent	
3	40.0	admin.	married	basic.6y	no	no	no	telephone	may	mon	151.0	1	999	0	nonexistent	
4	56.0	services	married	high.school	no	no	yes	telephone	may	mon	307.0	1	999	0	nonexistent	

Por otro lado, con *info(),* presenta la estructura/resumen del DataFrame, es decir, número de filas, tipos de datos, cantidad de valores nulos y memoria utilizada. Esto permite detectar variables categóricas, numéricas y posibles problemas de calidad de datos.

```
print("Información de columnas y tipos de datos:")
print(df.info())
```

[9] ✓ 0.0s

```
Información de columnas y tipos de datos:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 17 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             40612 non-null  float64
 1   job             39088 non-null  object
 2   marital         41188 non-null  object
 3   education       39912 non-null  object
 4   default         41188 non-null  object
 5   housing         41188 non-null  object
 6   loan            41188 non-null  object
 7   contact         41188 non-null  object
 8   month           41188 non-null  object
 9   day_of_week     41188 non-null  object
 10  duration        40324 non-null  float64
 11  campaign        41188 non-null  int64
 12  pdays           41188 non-null  int64
 13  previous        41188 non-null  int64
 14  poutcome        41188 non-null  object
 15  nr.employed     41188 non-null  float64
 16  y               41188 non-null  object
dtypes: float64(3), int64(3), object(11)
memory usage: 5.3+ MB
None
```

Posterior a eso, se calculó el porcentaje de valores nulos de cada columna con *isnull()* identificando los valores faltantes, luego con el *mean()* obtenemos la proporción y, esto, multiplicado por 100 convierte la proporción en porcentaje, con tal porcentaje decidimos si imputamos o eliminamos datos:

```
df.isnull().mean() * 100
```

[31] ✓ 0.0s

```
age               1.398466
job               5.098572
marital           0.000000
education         3.097990
default           0.000000
housing           0.000000
loan              0.000000
contact           0.000000
month             0.000000
day_of_week       0.000000
duration          2.097698
campaign          0.000000
pdays             0.000000
previous          0.000000
poutcome          0.000000
nr.employed       0.000000
y                 0.000000
dtype: float64
```

Eliminamos filas donde las variables *age* y *duration* tienen valores nulos debido a que son variables numéricas clave para el análisis, no las imputamos arbitrariamente, ya que afectaría el resultado tras el análisis, decidimos eliminarlas ya que no afecta demasiado si representa un porcentaje pequeño, 1.39% y 2.09%, respectivamente.

```
[32]  df = df.dropna(subset=['age', 'duration'])
      ✓ 0.0s
```

Tras hacer eso, generamos estadísticas descriptivas de las variables numéricas, con la finalidad de detectar outliers, ver rangos típicos y conocer la dispersión de los datos:

```
print("Estadísticas descriptivas - Variables Numéricas:")
df.describe()
✓ 0.0s
```

Estadísticas descriptivas - Variables Numéricas:

	age	duration	campaign	pdays	previous	nr.employed
count	39756.000000	39756.000000	39756.000000	39756.000000	39756.000000	39756.000000
mean	40.016174	258.592363	2.569423	962.484128	0.172905	5167.019630
std	10.415142	259.112332	2.775304	186.888325	0.494199	72.293521
min	17.000000	0.000000	1.000000	0.000000	0.000000	4963.600000
25%	32.000000	103.000000	1.000000	999.000000	0.000000	5099.100000
50%	38.000000	180.000000	2.000000	999.000000	0.000000	5191.000000
75%	47.000000	320.000000	3.000000	999.000000	0.000000	5228.100000
max	98.000000	4918.000000	56.000000	999.000000	6.000000	5228.100000

La línea *df.describe(include="object")* extrae estadísticas de variables categóricas, data necesaria para identificar si una categoría domina, ver distribución de valores e incluso detectar categorías poco frecuentes:

```
print("Estadísticas descriptivas - Variables Categóricas:")
df.describe(include='object')
✓ 0.0s
```

Estadísticas descriptivas - Variables Categóricas:

	job	marital	education	default	housing	loan	contact	month	day_of_week	poutcome	y
count	37742	39756	38529	39756	39756	39756	39756	39756	39756	39756	39756
unique	12	4	8	3	3	3	2	10	5	3	2
top	admin.	married	university.degree	no	yes	no	cellular	may	thu	nonexistent	no
freq	9508	24065	11365	31467	20864	32778	25245	13258	8322	34326	35269

2.3 Evaluación de la calidad de datos

2.3.1 Análisis de duplicados

Con la finalidad de saber cuántas filas del dataset están duplicadas completamente usamos el *df.duplicated().sum(),* seguido de un *df.drop_duplicated(inplace=true),* el cual elimina los registros duplicados directamente en el dataframe.

Realizamos este procedimiento debido a que es parte esencial para asegurar que el análisis no tenga sesgos.

```
    print("Análisis de Duplicados:")
    duplicados = df.duplicated().sum()
    print(f"   - Total de registros duplicados: {duplicados}")
    print(f"   - Porcentaje de duplicados: {(duplicados/len(df)*100):.2f}%")
```

✓ 0.0s

```
Análisis de Duplicados:
   - Total de registros duplicados: 11
   - Porcentaje de duplicados: 0.03%
```

```
    df.drop_duplicates(inplace=True)
```

✓ 0.0s

Observamos que había 11 duplicados y los eliminamos.

2.3.2 Análisis de valores "unknown"

El dataset no usa nulls en columnas categóricas, en su lugar, usa *"unknown"* que representaría un dato no reportado, cliente que no respondió esa pregunta o un registro incompleto en los sistemas del banco. Con la línea *df[df.isin(["unknown"]).any(axis=1)]* inspeccionamos dónde aparece ese valor.

```
df[df.isin(['unknown']).any(axis=1)]
```

✓ 0.0s

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	nr.employed	y
1	57.0	services	married	high.school	unknown	no	no	telephone	may	mon	149.0	1	999	0	nonexistent	5191.0	no
5	45.0	services	married	basic.9y	unknown	no	no	telephone	may	mon	198.0	1	999	0	nonexistent	5191.0	no
7	41.0	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	217.0	1	999	0	nonexistent	5191.0	no
10	41.0	blue-collar	married	unknown	unknown	no	no	telephone	may	mon	55.0	1	999	0	nonexistent	5191.0	no
15	54.0	retired	married	basic.9y	unknown	yes	yes	telephone	may	mon	174.0	1	999	0	nonexistent	5191.0	no
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
41118	34.0	technician	married	unknown	no	yes	no	cellular	nov	tue	162.0	2	999	2	failure	4963.6	no
41120	60.0	admin.	married	unknown	no	no	no	cellular	nov	tue	333.0	2	999	0	nonexistent	4963.6	no
41122	34.0	technician	married	unknown	no	no	no	cellular	nov	tue	985.0	3	999	0	nonexistent	4963.6	yes
41135	54.0	technician	married	unknown	no	yes	no	cellular	nov	thu	222.0	1	999	1	failure	4963.6	no
41175	34.0	student	single	unknown	no	yes	no	cellular	nov	thu	180.0	1	999	2	failure	4963.6	no

10301 rows × 17 columns

Luego de inspeccionar eso, utilizamos la línea *(df == 'unknown').sum()* para devolver cuántos "unknown" existen por columna, seguido de un *(df == 'unknown').mean() * 100* que equivaldría al porcentaje.

```
(df == 'unknown').mean()*100
✓ 0.0s
```

age	0.000000
job	0.774940
marital	0.201283
education	4.113725
default	20.845389
housing	2.405334
loan	2.405334
contact	0.000000
month	0.000000
day_of_week	0.000000
duration	0.000000
campaign	0.000000
pdays	0.000000
previous	0.000000
poutcome	0.000000
nr.employed	0.000000
y	0.000000

dtype: float64

La columna *default* tiene aproximadamente 21% de *"unknown",* esto indica problemas fuertes de calidad en esa variable.

2.3.3 Análisis de Outliers usando el método IQR

Aquí se evalúa todas las columnas numéricas, se genera una tabla con límites y porcentajes de outliers, con lo cual, podemos decidir si se eliminan, se capean, se transforman o se mantienen:

```
Análisis de Outliers (Variables Numéricas):
Usando el método del Rango Intercuartílico (IQR):
```

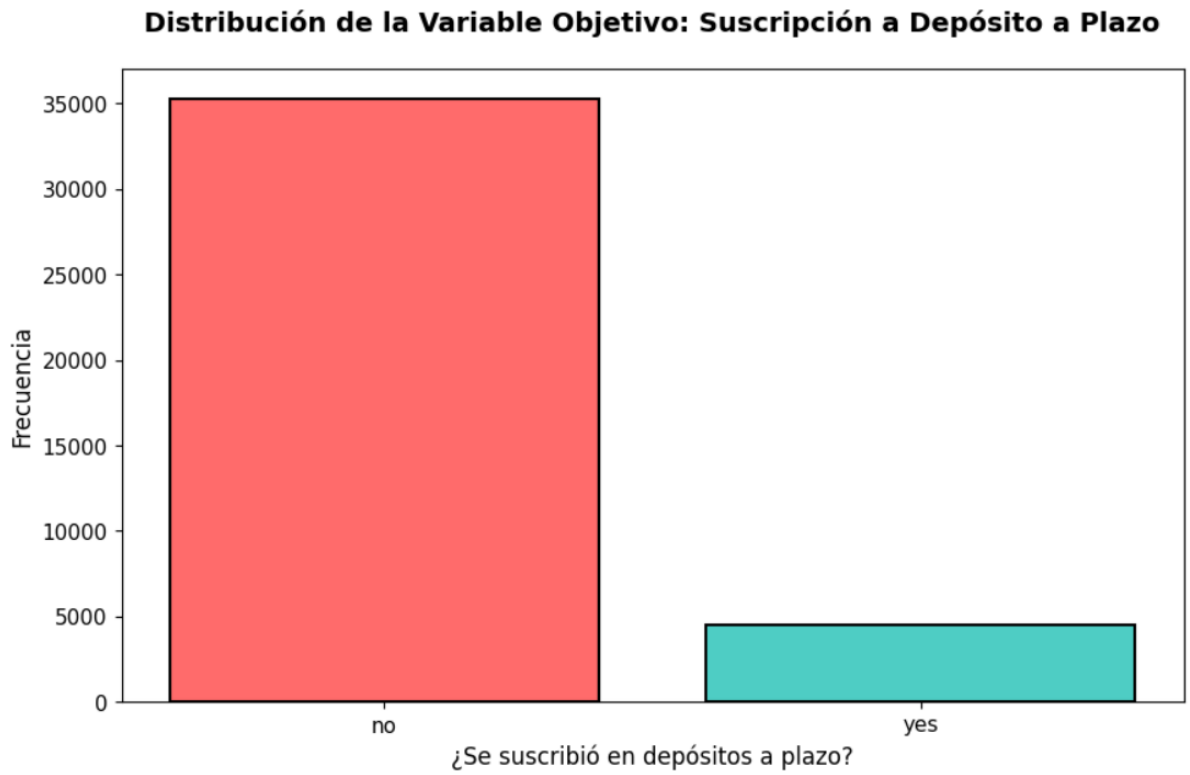
Variable	Q1	Q2	Q3	IQR	Límite Inferior	Límite Superior	Num. Outliers	Porcentaje
age	32.0	38.0	47.0	15.0	9.5	69.5	451	1.13%
duration	103.0	180.0	320.0	217.0	-222.5	645.5	2856	7.19%
campaign	1.0	2.0	3.0	2.0	-2.0	6.0	2335	5.87%
pdays	999.0	999.0	999.0	0.0	999.0	999.0	1462	3.68%
previous	0.0	0.0	0.0	0.0	0.0	0.0	5430	13.66%
nr.employed	5099.1	5191.0	5228.1	129.0	4905.6	5421.6	0	0.00%

3. Análisis de Distribución de Variables

3.1 Distribución de la Variable Objetivo (y)

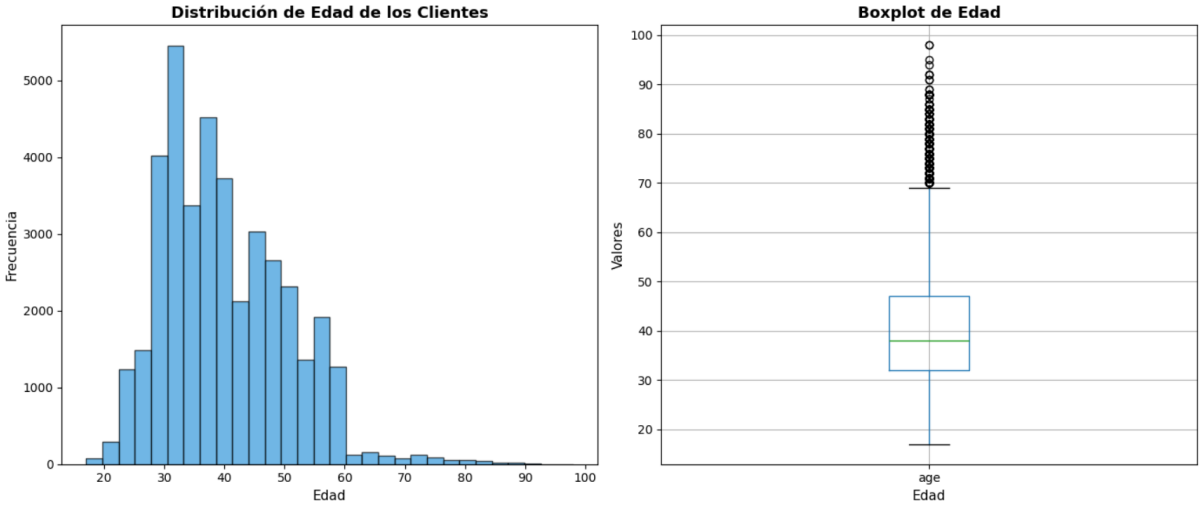
La variable *y* indica si el cliente aceptó un depósito a plazo, registra valores yes o no.

Con la gráfica podemos observar que la mayoría de los clientes no aceptaron la oferta, esto es algo que se intuye debido a que, en campañas bancarias, las tasas de conversión suelen ser bajas.



3.2. Distribución de la Edad

Para este análisis se usaron dos gráficos, un histograma para ver como se distribuye la variable *edad.* Y, el boxplot, para identificar visualmente valores mínimos, máximos, cuartiles y outliers.

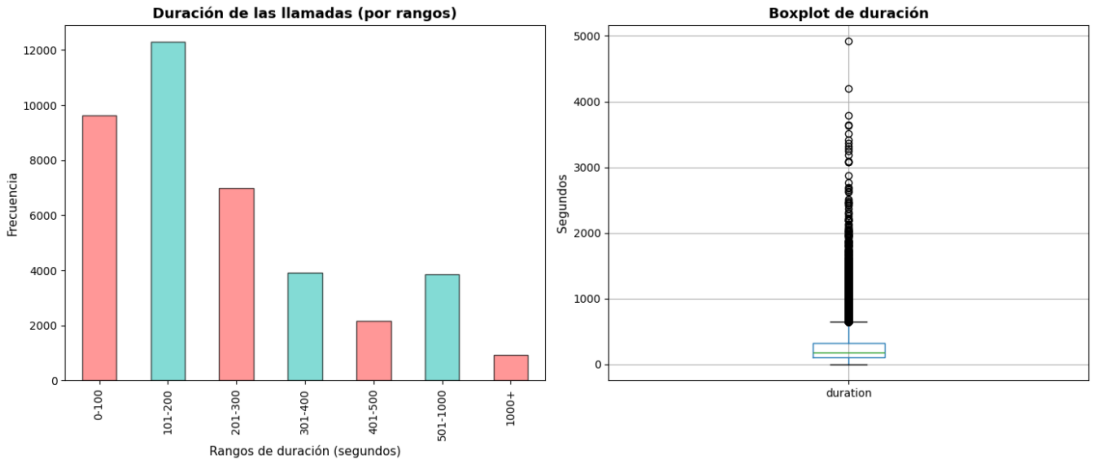


La mayoría de los clientes está entre 30 y 50 años, sin embargo, el boxplot, revela una enorme cantidad de outliers, los cuales corresponden a edades que superan los 69.5 años (límite superior RIQ). Los valores altos podrían ser de personas de edad avanzada, pero también indicar ruido o registros poco comunes.

3.3 Distribución de la Duración de las Llamadas

Con la gráfica podemos indicar que la mayoría de las llamadas duran hasta máximo 300 segundos (5 min), a su vez, que hay valores extremos que superan 645 segundos (10 min), las llamadas muy largas suelen estar asociadas a clientes que aceptaron la oferta.

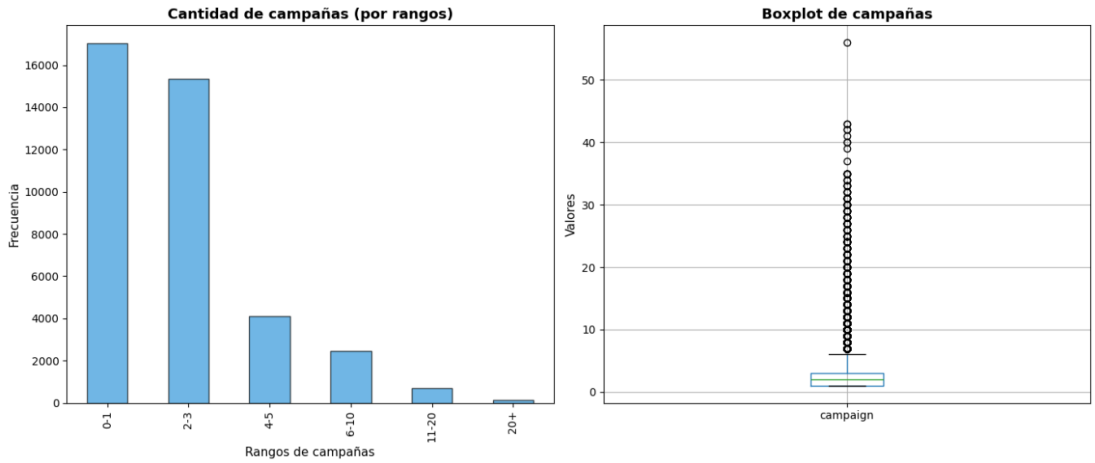
Este comportamiento es habitual en marketing telefónico, debido a que llamadas largas tienden a ser exitosas.



3.4 Distribución de la Cantidad de Contactos por Campaña

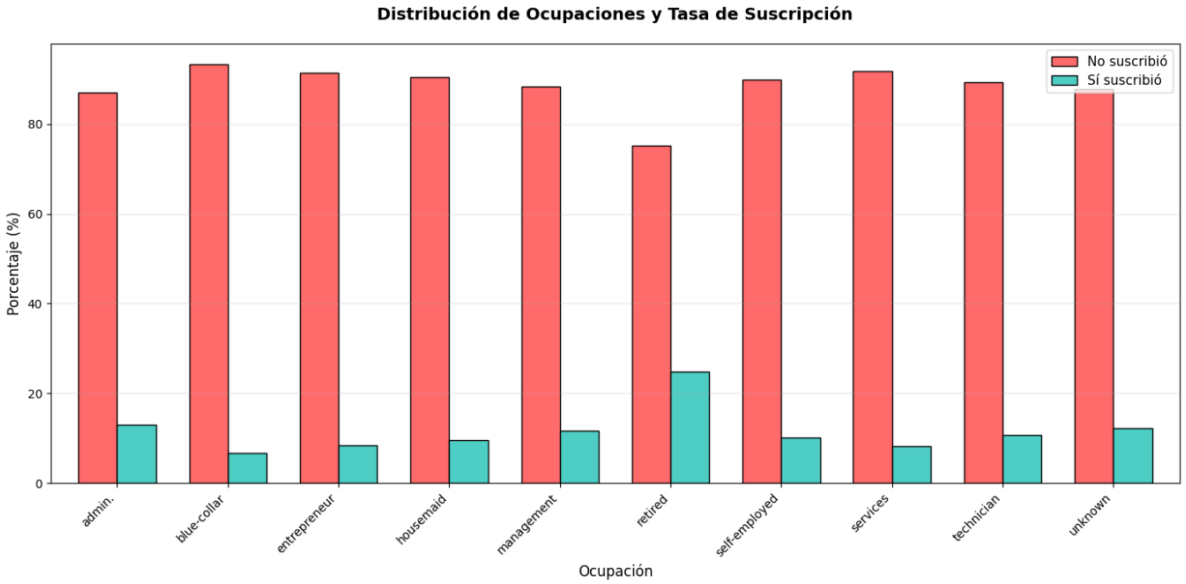
La mayor parte de los clientes fueron contactados entre 0 y 3 veces. Estos dos primeros rangos concentran más de la mitad del dataset. A partir de 4 contactos, la frecuencia empieza a caer drásticamente. Existen casos extremos donde el cliente fue contactado más de 20 veces, pero son muy pocos.

La variable campaign tiene una distribución muy sesgada: hay muchísimos valores bajos y muchos outliers, lo cual indica que un pequeño grupo de clientes fue contactado de forma excesiva.



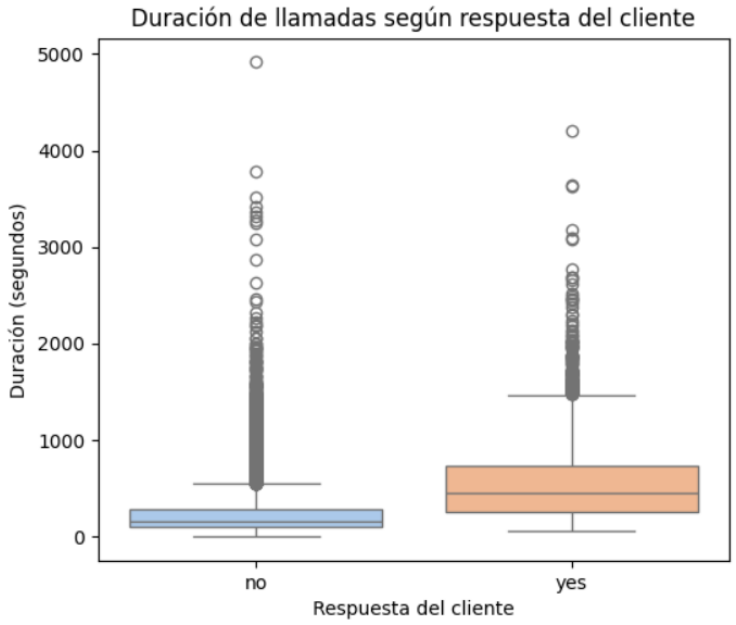
3.5 Distribución de Ocupaciones y Probabilidad de Suscripción

Este gráfico compara, por ocupación el porcentaje de quien sí y quien no se suscribió. Gracias a la gráfica, podemos indicar que el grupo *retired* es el que más se suscribe. Esto puede deberse a que suelen tener más capital disponible, buscan productos de ahorro e inversión y son más receptivos a depósitos a plazo.



3.6 Duración de llamadas según respuesta (y)

Las llamadas donde el cliente dijo “no” son más cortas, al contrario de las llamadas donde el cliente dijo “sí”, aquí las llamadas son significativamente más largas, Este es uno de los predictores más fuertes en el dataset.

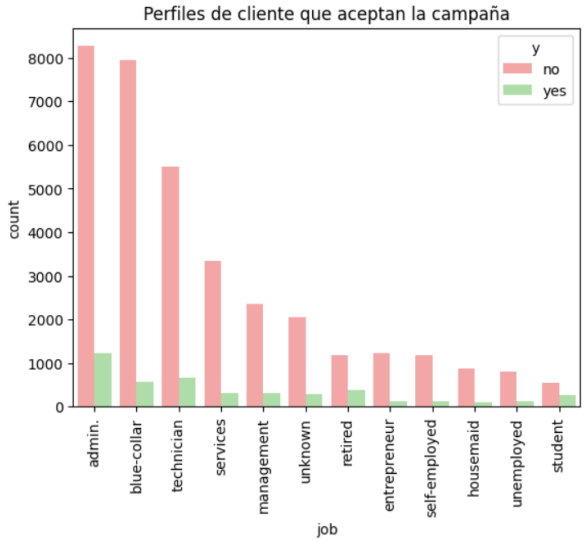


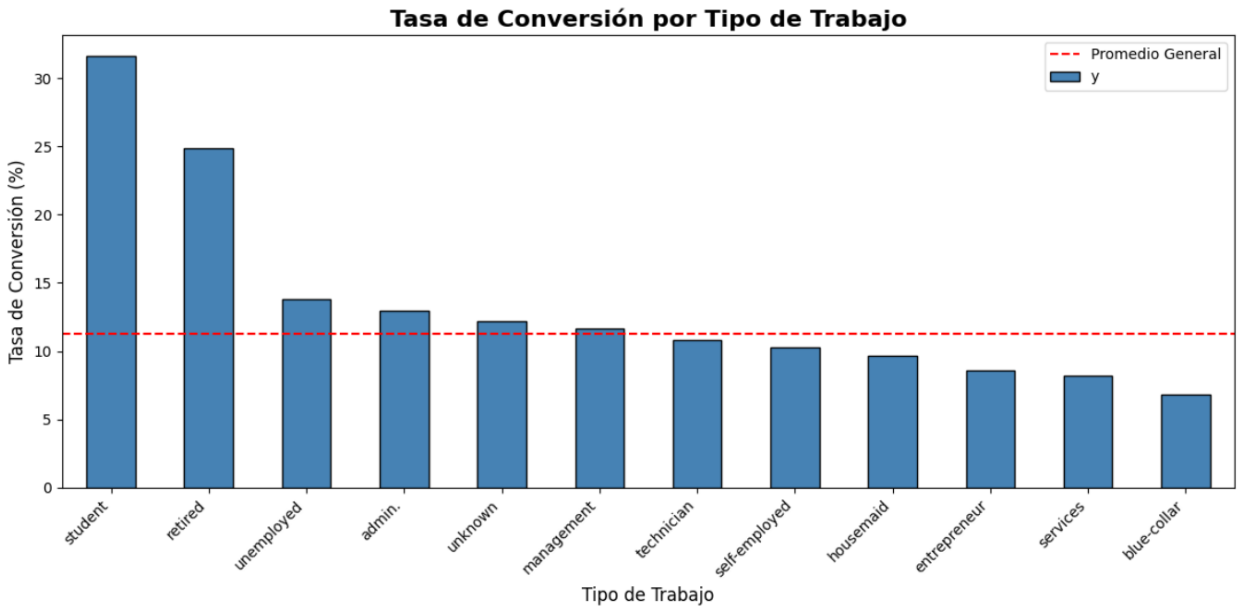
4. Insights del Proyecto

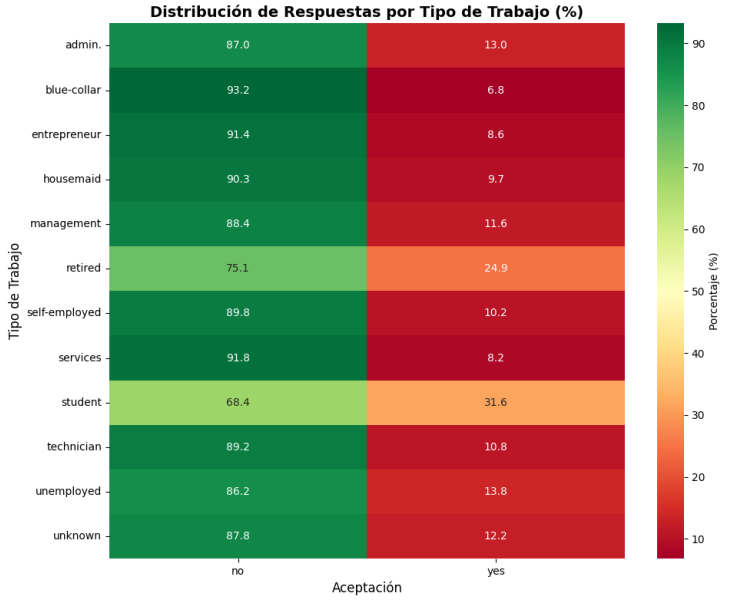
4.1 Perfiles de clientes que aceptan la campaña

Al analizar la variable job (ocupación), se identificó que la distribución de contactos no es proporcional a la tasa de aceptación de los clientes. Aunque los grupos ocupacionales como *admin* y *blue-collar* concentran la mayor cantidad de llamadas realizadas, estos no son necesariamente los más receptivos.

Al profundizar con una visualización de la tasa de aceptación por tipo de trabajo, se evidencia que existen ocupaciones con poblaciones pequeñas que presentan tasas de respuesta positiva más elevadas. Este hallazgo sugiere una oportunidad: redirigir parte del esfuerzo comercial hacia estos segmentos menos contactados pero más propensos a aceptar.





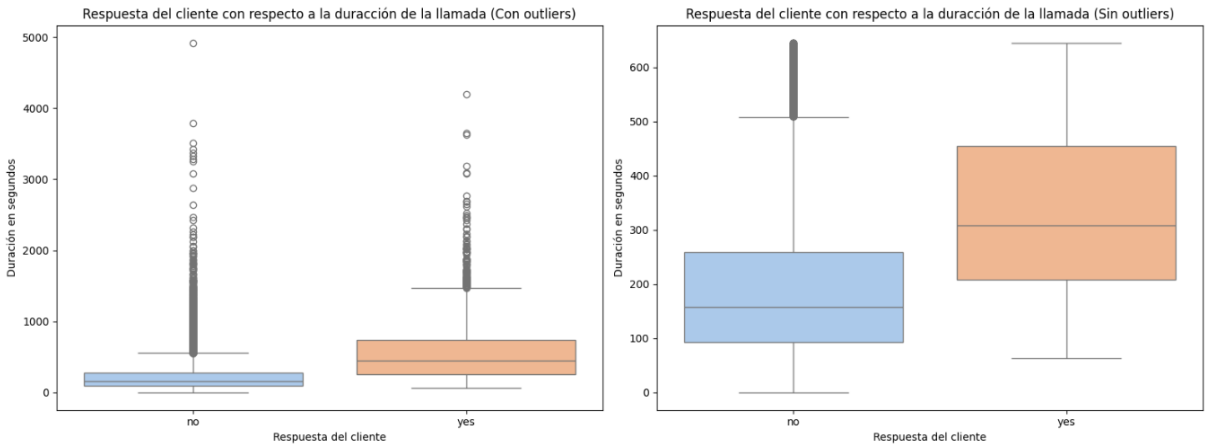


En términos estratégicos, esto invita a equilibrar la distribución de llamadas por ocupación y validar si estos patrones se mantienen a lo largo del tiempo. De confirmarse, podrían convertirse en segmentos prioritarios para la próxima campaña.

4.2 Influencia de la duración de la llamada en la aceptación del cliente

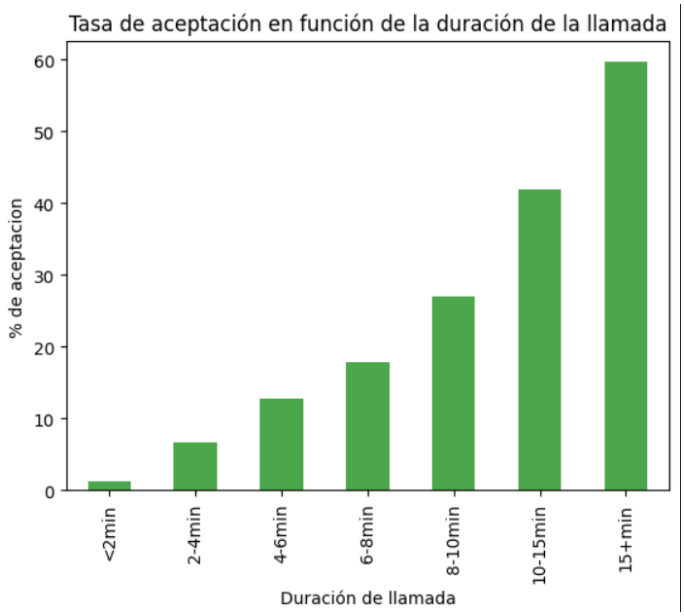
La variable duration (duración de la llamada) mostró una relación directa con la probabilidad de aceptación. En términos simples, mientras más tiempo dure la llamada, mayor es la probabilidad de obtener un "sí".

Al comparar los gráficos con outliers y sin outliers, se observa claramente que las llamadas asociadas a respuestas positivas tienden a ser más largas.



Incluso al categorizar estas duraciones por rangos (por ejemplo, 2–4 minutos, 10–15 minutos, etc.), se confirma que los clientes que aceptaron el producto permanecieron más tiempo en la conversación.

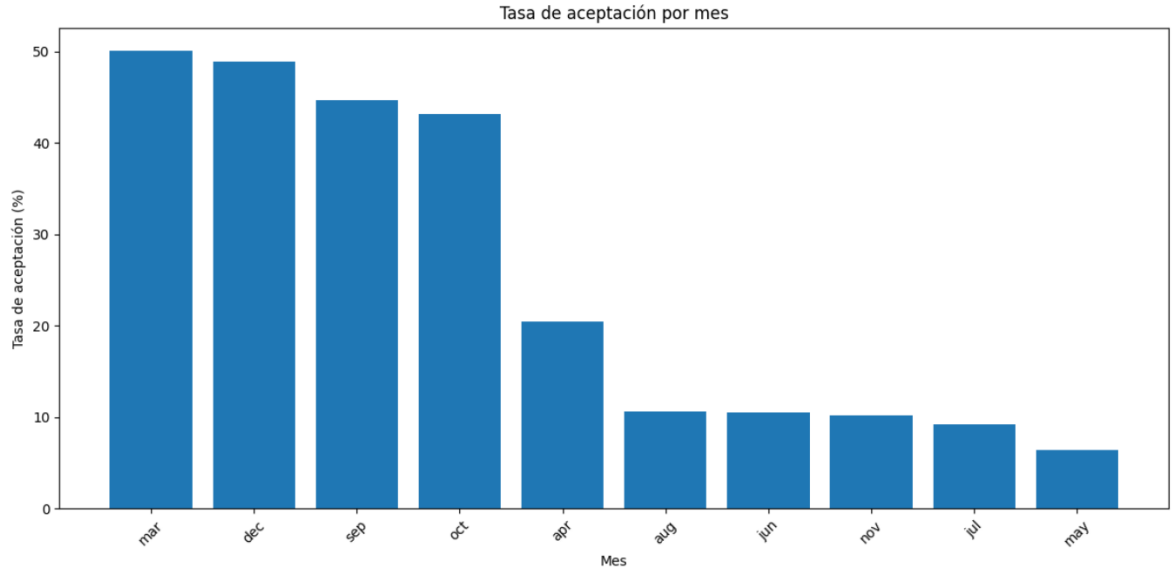
No obstante, esta relación no es determinante. A pesar de que muchas llamadas largas terminan en aceptación, también existen llamadas prolongadas que resultan en un “no”. Esto refleja un comportamiento importante: una llamada larga indica interés, pero no garantiza conversión.



Operativamente, este hallazgo destaca la necesidad de fortalecer las competencias del equipo de call center para que puedan identificar cuándo un cliente está realmente interesado y cuándo no, evitando inversiones innecesarias de tiempo en llamadas poco prometedoras.

4.3 Meses con mayor probabilidad de aceptación

El análisis temporal evidenció que existen meses del año donde la probabilidad de aceptación es considerablemente más alta. Entre estos destacan: marzo, diciembre, septiembre y octubre, meses donde la tasa de aceptación supera el 40%.

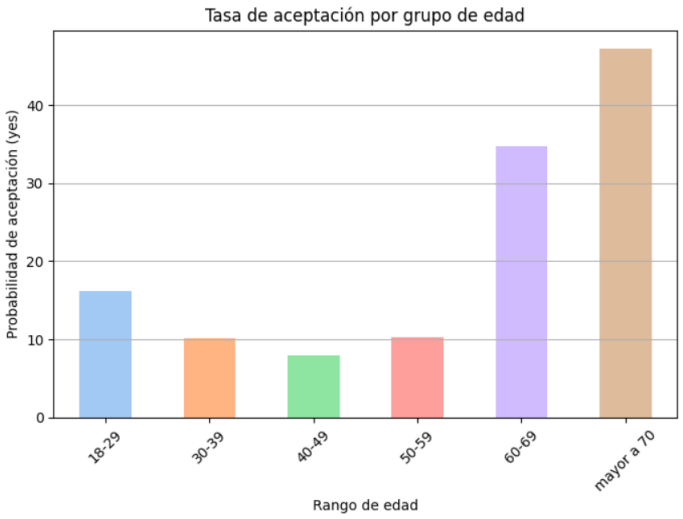


Un aspecto importante es que, en algunos casos, estos meses muestran tasas elevadas a pesar de tener volúmenes de contacto relativamente bajos. Esto implica que podrían estar subutilizados. Incrementar el número de llamadas en estos meses podría mejorar significativamente los resultados globales sin necesidad de modificar la propuesta comercial.

Este patrón estacional sugiere la importancia de planificar campañas específicas en meses de alta receptividad para maximizar la eficiencia operativa.

4.4 Rangos de edad con mayor probabilidad de éxito

Al segmentar la edad de los usuarios en grupos, se observó que tanto los clientes más jóvenes (18–29 años) como las personas mayores de 60 años, especialmente los mayores de 70, presentan tasas de aceptación superiores a otros grupos.



Estos hallazgos sugieren que las decisiones financieras están influenciadas de manera diferente según el ciclo de vida del cliente.

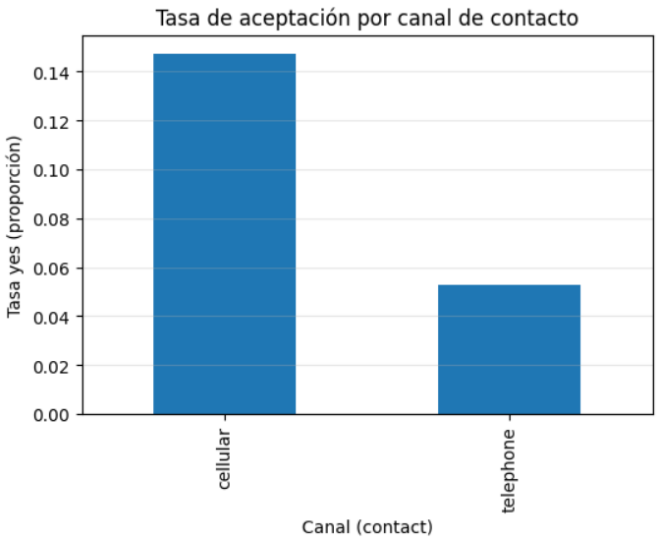
Los más jóvenes pueden estar más abiertos a productos nuevos, mientras que los adultos mayores podrían tener necesidades específicas relacionadas con ahorro, inversión o estabilidad financiera.

Este comportamiento abre la puerta a desarrollar campañas personalizadas para estos segmentos, considerando su lenguaje, motivaciones y necesidades particulares.

4.5 El canal telefónico como medio de contacto presenta baja aceptación

El análisis del canal utilizado para contactar a los clientes evidencia que la vía telefónica tiene la tasa de aceptación más baja frente a otros canales disponibles.

Este indicador plantea varias hipótesis: el cliente podría percibir las llamadas como invasivas, no siempre es el momento oportuno o, existe saturación por parte de otras campañas telefónicas.



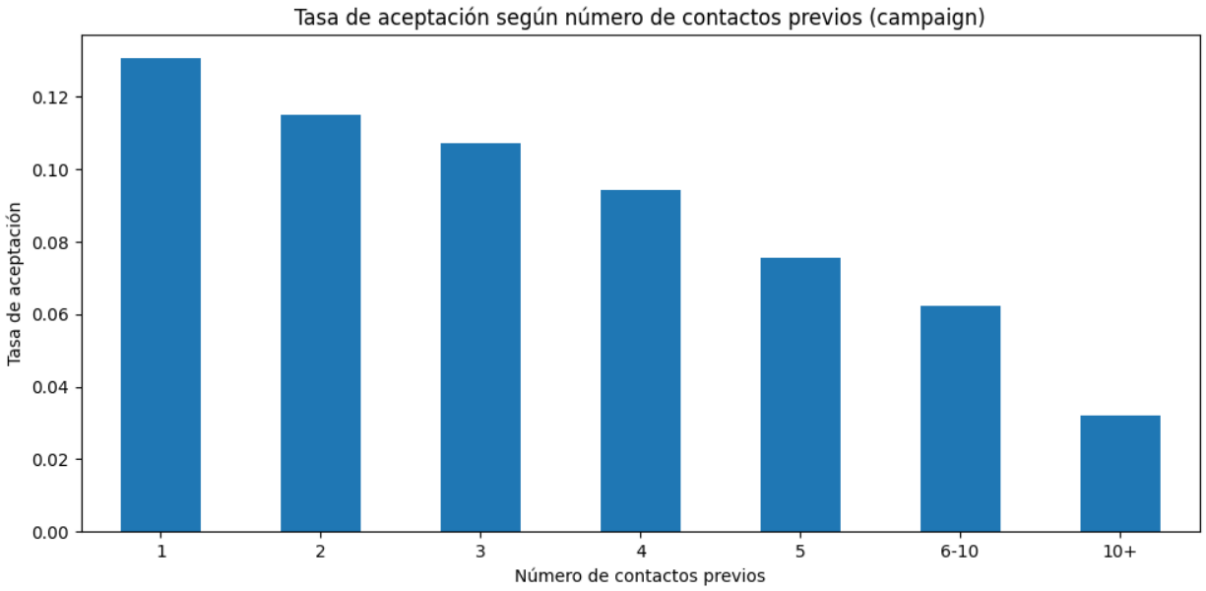
Este resultado invita a evaluar canales alternativos o combinar la llamada con un contacto previo por otros medios (correo, SMS o notificaciones), para reducir la fricción inicial.

4.6 Mayor número de contactos previos reduce la probabilidad de aceptación

Finalmente, el análisis de las variables poutcome (resultado de campañas anteriores) y campaign (número de contactos durante la campaña actual) muestra que la persistencia excesiva reduce la probabilidad de aceptación.

Cuando una persona es contactada repetidamente dentro de la misma campaña, la tasa de conversión disminuye, e incluso se observan respuestas negativas constantes en clientes con intentos previos.

Este comportamiento sugiere la necesidad de definir un límite óptimo de contactos para evitar que la insistencia genere rechazo.



5. Conclusiones

El análisis realizado permitió obtener una visión integral del comportamiento de los clientes ante las campañas de marketing del banco, identificando patrones clave que pueden guiar la toma de decisiones en futuras estrategias comerciales. Uno de los hallazgos más relevantes fue la notable presencia de problemas de calidad de datos, como valores duplicados, registros incompletos y una proporción significativa de categorías "unknown", especialmente en variables sensibles como *default*. Abordar estas inconsistencias permitió asegurar un análisis más confiable y establecer un punto de partida sólido para la exploración posterior.En cuanto al análisis descriptivo, las variables demográficas y de comportamiento mostraron tendencias claras. La edad presentó una distribución amplia con un número considerable de valores atípicos, confirmando que la base de clientes abarca un rango etario diverso. Por su parte, la duración de las llamadas emergió como uno de los predictores más influyentes: se observó que las conversaciones asociadas a respuestas positivas suelen ser sustancialmente más largas, lo cual refuerza su importancia operativa dentro de las campañas.

Asimismo, se evidenció que la ocupación tiene un papel relevante en la probabilidad de aceptación. Aunque algunos grupos concentran la mayor parte de los contactos, no son necesariamente los más receptivos. Ocupaciones como *retired* muestran una tendencia mucho mayor a aceptar el producto, lo que revela la existencia de segmentos altamente valiosos que podrían estar subatendidos por las estrategias actuales.

El análisis temporal añadió una perspectiva estratégica complementaria. Meses como marzo, septiembre, octubre y diciembre presentaron tasas de aceptación superiores al promedio, incluso cuando el volumen de contactos no fue el más alto. Este hallazgo sugiere que existe una estacionalidad favorable que puede ser aprovechada para optimizar el rendimiento de las campañas sin incrementar de manera significativa los costos operativos.

No obstante, también se identificaron limitaciones en el enfoque actual de contacto. El canal telefónico mostró una baja tasa de conversión y la insistencia excesiva, evidenciada por altos valores en la variable *campaig,* pareció tener un efecto contraproducente. Esto resalta la necesidad de replantear el uso del canal telefónico y de establecer límites óptimos de contacto para evitar rechazo por saturación.

En conjunto, los hallazgos obtenidos permiten concluir que la aceptación de un depósito a plazo depende de una combinación de factores demográficos, temporales y conductuales. El estudio ofrece una base sólida para diseñar campañas más focalizadas, mejorar los procesos de

segmentación e incrementar la eficiencia en la gestión comercial. Optimizar la asignación de recursos, respetar los límites de contacto y priorizar segmentos con alta probabilidad de aceptación se perfilan como acciones clave para elevar los resultados en futuras campañas del banco.