Caracterización de Clientes Bancarios y Comportamientos de Campaña mediante Análisis Exploratorio y Técnicas de Preprocesamiento

Characterization of Banking Customers and Campaign Behaviors through Exploratory Analysis and Preprocessing Techniques

Geraldine Acevedo Restrepo
Facultad de Ingeniería, Universidad de Antioquia

Resumen

El análisis realizado sobre el Bank Marketing Dataset permitió comprender mejor el perfil y comportamiento de los clientes contactados por el banco. En general, la mayoría son adultos entre los 30 y 50 años, con niveles educativos de secundaria o universidad, y una proporción importante está casada y tiene préstamo de vivienda. Las variables numéricas mostraron patrones muy distintos: la edad fue estable, mientras que el saldo presentó una gran variación, desde valores muy bajos hasta montos extremadamente altos o negativos. La duración de las llamadas también fue muy desigual, con la mayoría de conversaciones cortas y algunas pocas excepcionalmente largas.

Al comparar variables entre sí, no se encontraron relaciones lineales fuertes; cada variable aporta información independiente. Las asociaciones entre variables categóricas fueron significativas, pero muy débiles, por lo que no afectan de manera importante el análisis general. Los valores atípicos identificados en balance y duration no representan errores, sino comportamientos reales dentro del contexto bancario.

Por último, el preprocesamiento permitió dejar el dataset en mejores condiciones. Age solo requirió estandarización, mientras que balance y duration necesitaron transformaciones más profundas para corregir su fuerte asimetría y preparar los datos para análisis futuros.

Palabras clave: análisis exploratorio, preprocesamiento, outliers, marketing bancario, transformacioness.

1. Introducción

Entender cómo se comportan los clientes de un banco es clave para diseñar buenas campañas de marketing, mejorar la atención y saber qué tipo de usuarios podrían aceptar una oferta. Los análisis exploratorios ayudan a identificar características importantes como la edad, el nivel educativo, el estado civil, los saldos bancarios y la forma en la que responden a una llamada comercial.

En este trabajo realice el análisis del Bank Marketing Dataset, un conjunto de datos obtenido de la plataforma Kaggle (Moro et al., 2014). El objetivo principal es describir cómo se comportan las variables seleccionadas, revisar si existen patrones entre ellas y dejar el dataset preparado para etapas futuras, como podría ser el uso de modelos predictivos.

Para lograrlo se realizaron análisis univariados, bivariados y multivariados, además de una revisión de valores atípicos y un preprocesamiento básico. Esto permitió entender mejor el tipo de clientes que aparecen en el conjunto de datos y cómo podrían relacionarse entre sí.

2. Metodología

2.1 Dataset y selección de variables

Para este análisis se eligieron seis variables que representan aspectos importantes de los clientes: edad, saldo promedio, duración de la llamada, estado civil, nivel educativo y si tienen o no un préstamo de vivienda.

Las variables numéricas permiten ver cómo se comportan características financieras y demográficas, mientras que las categóricas ayudan a entender el contexto social y económico del cliente. En conjunto, estas variables dan una visión equilibrada del tipo de personas contactadas por el banco.

2.2 Análisis univariado

El análisis univariado consistió en revisar cada variable por separado para identificar su comportamiento general. En el caso de las variables numéricas se calcularon medidas básicas como la media, la mediana, los rangos y la desviación estándar. También se evaluó la forma de sus distribuciones y si tenían valores extremos. Esto se complementó con gráficos como histogramas y boxplots

que permitieron confirmar visualmente lo que mostraban los números.

La variable age mostró una distribución estable, con la mayoría de los clientes entre 30 y 50 años, aunque existen algunos casos aislados de personas mayores de 70 años. Balance fue la variable más irregular, la mayoría de los clientes tienen saldos bajos, pero existen casos con saldos muy altos que generan una cola larga en la distribución. Duration también presentó una fuerte asimetría, ya que casi todas las llamadas fueron cortas, pero hubo algunas que se extendieron bastante, incluso hasta más de 80 minutos.

Las variables categóricas se revisaron usando tablas de frecuencia. Allí se observó que la mayoría de los clientes están casados, que los niveles educativos más comunes son secundaria y terciaria, y que más de la mitad tiene un préstamo de vivienda. Estas distribuciones son coherentes con una población adulta y activa laboralmente.

Al final los cálculos de asimetría y curtosis confirmaron que balance y duration no siguen una distribución normal debido a sus valores extremos, mientras que age es mucho más estable. Esto indica que las dos primeras variables podrían necesitar transformaciones en etapas posteriores, mientras que age puede utilizarse tal como está.

3. Análisis Bivariado y Multivariado

Para analizar las correlaciones, distribuciones y asociaciones entre variables se emplearon herramientas de visualización y estadística implementadas mediante Pandas y Seaborn (McKinney, 2010; Waskom, 2021). El primer paso consistió en examinar la correlación entre las variables numéricas age, balance y duration. La matriz de correlaciones reveló coeficientes muy bajos en todos los pares, con valores absolutos

inferiores a 0.1, lo que evidencia ausencia de relaciones lineales fuertes. Esto fue confirmado visualmente mediante scatterplots, donde las nubes de puntos no mostraron tendencias definidas. Estos resultados sugieren que cada variable aporta información independiente, lo que elimina posibles problemas de colinealidad y permite un uso conjunto en modelos posteriores sin redundancia.

Debido al gran tamaño del dataset, algunas visualizaciones se realizaron sobre una muestra aleatoria del 20 %, manteniendo las proporciones originales y preservando la estructura general de las distribuciones. Esta estrategia permitió obtener gráficos más legibles sin comprometer la representatividad estadística. Las conclusiones se mantuvieron consistentes tanto con la muestra como con el dataset completo: ni la edad explica el saldo, ni el saldo condiciona la duración de la llamada, ni existe relación entre la edad y la extensión de la conversación telefónica.

El análisis de relaciones entre variables continuas y categóricas se llevó a cabo mediante boxplots y violin plots. Estos gráficos permitieron comparar la distribución de balance y duration entre diferentes categorías de housing, marital y education.

En todos los casos se observó una tendencia común y es que las medianas de los grupos eran similares y las diferencias más notorias provenían de la dispersión y de las colas largas asociadas a los outliers. Por ejemplo, las personas con préstamo de vivienda presentaron mayor variabilidad en los saldos, pero sin diferencias relevantes en la mediana; y los niveles educativos mostraron patrones casi idénticos, con la mayoría de los clientes concentrados cerca de saldos bajos. La duración de la llamada tampoco presentó cambios significativos según el estado civil.

Estos resultados confirman que las variables categóricas examinadas no modifican de manera sustancial las distribuciones de las variables continuas.

Para analizar relaciones entre variables categóricas se utilizaron tablas cruzadas y pruebas Chi-cuadrado estas se encuentran implementadas en SciPy (Virtanen et al., 2020). Los resultados indicaron asociaciones estadísticamente significativas en todos los casos, pero de magnitud muy débil, como lo mostraron los valores de V de Cramer, siempre inferiores a 0.13. Esto significa que, aunque el tamaño de la muestra permite detectar diferencias, en términos prácticos las relaciones son mínimas. Aun así, los residuales estandarizados permiten observar patrones coherentes como que los solteros tienden a no tener préstamo de vivienda, los casados muestran mayor presencia de créditos hipotecarios, y los niveles educativos se distribuyen de manera distinta entre estados civiles, reflejando posibles diferencias generacionales. Estos patrones son interesantes desde una perspectiva descriptiva, pero no representan relaciones suficientemente fuertes como para influir de manera determinante en el comportamiento general del dataset.

Teniendo en cuenta tanto el análisis bivariado y multivariado se demuestra que las variables seleccionadas operan de forma predominantemente independiente, con asociaciones leves pero coherentes entre algunas variables categóricas. Esta independencia refuerza la utilidad de cada variable como fuente de información complementaria y facilita su integración en etapas de modelado sin riesgos de duplicar información o generar multicolinealidad.

4. Detección de Valores Atípicos (Outliers)

La detección de valores atípicos

fue importante porque este tipo de datos puede influir notablemente en los resultados del análisis. En este conjunto, las variables balance y duration mostraron desde el principio comportamientos inusuales, como saldos muy altos o llamadas excepcionalmente largas.

Se usó el método del rango intercuartílico (IQR) para identificar valores extremos. En balance se encontró que aproximadamente el 5 % de los clientes tiene saldos muy por encima o por debajo del rango típico. Los saldos negativos representan posibles deudas, mientras que los saldos muy altos corresponden a clientes con una capacidad económica alta.

En duration, solo alrededor del 2.5 % de las llamadas fueron extremadamente largas, algunas superiores a 80 minutos. Estos casos no son errores, sino situaciones reales, aunque poco comunes.

No se aplicaron métodos como Z-Score o el test de Grubbs porque estos requieren distribuciones normales, algo que claramente no ocurre en este dataset.

Para complementar el análisis se usaron métodos multivariados como DBSCAN e Isolation Forest estás usan las implementaciones disponibles en Scikit-Learn (Pedregosa et al., 2011), que permiten detectar outliers considerando varias variables al mismo tiempo. Ambos métodos identificaron un pequeño grupo de clientes con combinaciones inusuales de edad, saldo y duración, confirmando la existencia de comportamientos atípicos pero válidos.

Todos estos valores extremos forman parte natural del conjunto de datos y no deben eliminarse, sino manejarse con técnicas de transformación o escalado cuando se utilicen en modelos predictivos. Su presencia aporta información útil sobre clientes con comportamientos poco frecuentes.

5. Imputación, escalamiento y transformación de datos

Antes de realizar algún tipo de imputación o transformación, fue necesario revisar si el dataset presentaba problemas como valores faltantes, escalas muy diferentes entre las variables o distribuciones demasiado desiguales que pudieran afectar el análisis. Lo que sucedió fue que al hacer la verificación inicial, se encontró que ninguna de las variables contenía datos nulos. Esto permitió avanzar directamente a la etapa de escalamiento que emplea herramientas como StandardScaler, RobustScaler, MinMaxScaler y la transformación Yeo–Johnson, implementados también en Scikit-Learn (Pedregosa et al., 2011) y transformación sin necesidad de imputar valores, lo cual facilita el proceso y garantiza que los resultados no dependan de suposiciones adicionales.

Una vez confirmada la ausencia de valores faltantes, se analizaron las características propias de cada variable para decidir qué tipo de transformación aplicar. En este caso, age mostró una distribución bastante estable y sin grandes desviaciones, por lo que únicamente requería una estandarización para trabajar en la misma escala que el resto. Balance, por el contrario, presentaba una dispersión muy amplia, con tanto valores extremadamente bajos como depósitos muy altos, lo que hacía necesario un método de escalamiento más robusto que no se viera afectado por estas variaciones. Lo que pasaba con duration era que claramente era una variable muy sesgada hacia la derecha, pues la mayoría de las llamadas eran cortas, pero también existían algunas que duraban varios minutos más de lo habitual, incluso hasta más de una hora. Por esta razón, también requería un tratamiento más profundo para suavizar su distribución.

Antes de realizar el escalamiento, se revisaron los histogramas originales de las tres variables. Esto permitió observar de forma clara la diferencia entre ellas, age tenía una forma relativamente normal, mientras que balance y duration mostraban

distribuciones muy asimétricas, confirmando la necesidad de transformarlas. Revisar estas gráficas antes de aplicar cualquier modificación permitió justificar las técnicas elegidas y tener un punto de comparación para el “antes y después”.

Luego se aplicaron los métodos de escalamiento. Para age se utilizó el escalador estándar, que centra la variable en cero y ajusta su desviación estándar a uno. Esto no cambia la forma de la distribución, pero facilita el análisis cuando las variables deben trabajar juntas. A balance se le aplicó un escalador robusto, ideal para situaciones donde existen valores extremos que pueden distorsionar la escala general. En el caso de duration, se optó por MinMaxScaler, que transforma los valores al rango entre 0 y 1. Este método es útil en modelos donde las magnitudes afectan directamente los cálculos, como los que dependen de distancias o normalización estricta.

Además del escalamiento, se aplicó una transformación Yeo–Johnson a balance y duration, debido a sus distribuciones altamente sesgadas. Esta técnica es especialmente conveniente porque funciona incluso cuando los datos presentan valores negativos o ceros. En balance, la transformación redujo la influencia de los valores extremadamente altos y acercó la distribución a una forma más manejable. En duration, el cambio fue especialmente notable si vemos la forma original tenía una cola muy larga hacia la derecha, se transformó en una distribución mucho más suave, equilibrada y fácil de interpretar.

Después de realizar las transformaciones y escalamientos, se graficaron nuevamente las variables para comparar los resultados. En balance se observó que la forma seguía mostrando variación, pero ya no era tan extrema, y los valores altos dejaban de dominar visualmente el histograma. En duration, la diferencia fue aún más evidente: la

distribución se volvió mucho más simétrica, perdiendo gran parte del sesgo original y mostrando una forma más parecida a una curva regular. Por su parte, age quedó correctamente estandarizada, lista para integrarse con las demás variables sin generar desequilibrios debido a diferencias en sus escalas.

Por lo que podemos decir que la falta de valores faltantes simplificó el proceso de preparación de los datos, y el escalamiento junto con las transformaciones seleccionadas permitió que las variables quedarán en un estado adecuado para análisis posteriores. Las transformaciones aplicadas a balance y duration fueron especialmente importantes, ya que ayudaron a corregir las fuertes asimetrías iniciales, mientras que age solo necesitó una estandarización sencilla.

6. Resultados

Los análisis realizados permitieron obtener una visión clara y completa del comportamiento de los clientes dentro del subconjunto del Bank Marketing Dataset. En todo el análisis se encontró que las variables seleccionadas describen adecuadamente aspectos sociales, demográficos, económicos y de interacción, lo que ayuda a comprender mejor el perfil de los usuarios contactados por el banco.

En primer lugar, el análisis univariado mostró que la mayor parte de los clientes son adultos entre los 30 y 50 años, un rango típico de personas económicamente activas. La variable balance reveló una fuerte desigualdad pues mientras la mayoría de los clientes tiene saldos bajos, existe un pequeño grupo con valores extremadamente altos o negativos. Esto confirma que la distribución financiera del conjunto es muy heterogénea. En cuanto a duration, la mayoría de las llamadas fueron breves, aunque existieron algunos casos aislados de conversaciones muy largas, lo que explica la fuerte asimetría observada. Las

variables categóricas, por su parte, mostraron distribuciones coherentes pues predominan los clientes casados, con niveles educativos secundarios o terciarios, y un porcentaje importante tiene crédito de vivienda.

El análisis bivariado permitió determinar que las variables numéricas no tienen relaciones lineales fuertes entre sí. La edad no explica el saldo, el saldo no explica la duración de la llamada y cada variable parece aportar información independiente. Cuando se compararon variables numéricas con categóricas, se encontró que las medianas de los grupos eran parecidas y que las diferencias estaban más asociadas a la variabilidad interna y a los valores extremos que a diferencias reales entre categorías. En las relaciones categóricas, aunque las pruebas estadísticas indicaron asociaciones significativas, estas fueron muy débiles. Los patrones encontrados como por ejemplo que los solteros suelen no tener vivienda propia o que cierta educación está más presente en ciertos estados civiles son coherentes, pero no lo suficientemente fuertes como para afirmar que influyen decisivamente en el comportamiento analizado.

En cuanto a los valores atípicos, se confirmó que balance y duration contienen observaciones extremas tanto por arriba como por abajo. Sin embargo, estos valores no deben considerarse errores, sino casos reales que forman parte natural del tipo de información que recopila un banco. Los métodos aplicados permitieron identificar alrededor del 5 % de casos extremos en balance y un poco más del 2 % en duration. Métodos más avanzados como DBSCAN e Isolation Forest también encontraron un pequeño grupo de clientes con combinaciones poco comunes de edad, saldo o duración de llamada. Esto refuerza la idea de que el dataset contiene un segmento reducido de clientes “atípicos”, pero válidos, que deben ser tratados adecuadamente y no eliminados.

Finalmente, la etapa de escalamiento y transformación permitió dejar las variables listas para análisis futuros. Balance y duration necesitaron transformaciones más profundas porque sus distribuciones originales eran muy irregulares, con colas largas y valores extremos. Las transformaciones aplicadas lograron reducir estos efectos, haciendo que las variables fueran más estables y fáciles de manejar. Age, al ser más regular, solo requirió una estandarización simple.

7. Conclusiones

1. El análisis permitió conocer con claridad el perfil general de los clientes del banco: la mayoría son adultos entre 30 y 50 años, con niveles educativos medios o altos y una participación importante de personas casadas y con préstamo de vivienda.

2. Las variables numéricas mostraron comportamientos muy distintos entre sí. Age fue estable y sin grandes variaciones, mientras que balance y duration presentaron distribuciones muy irregulares con valores extremos que afectan su forma y dispersión.

3. No se encontraron relaciones fuertes entre las variables numéricas. Ni la edad explica el saldo, ni el saldo explica la duración de la llamada, lo que indica que cada una aporta información independiente.

4. Las relaciones entre variables categóricas sí resultaron estadísticamente significativas, pero de forma muy débil, por lo que no influyen de manera importante en el comportamiento general del dataset.

5. Los valores atípicos identificados en balance y duration no deben considerarse errores, pues corresponden a casos reales de clientes con condiciones poco comunes. Estos deben tratarse adecuadamente y no eliminarse.

6. Los métodos multivariados (DBSCAN e Isolation Forest) confirmaron que existe

un pequeño grupo de clientes con combinaciones inusuales de características, lo que coincide con los patrones observados en el análisis univariado.

7. El proceso de escalamiento y transformación fue fundamental para dejar el dataset en mejores condiciones. Balance y duration necesitaban ajustes más profundos debido a su fuerte asimetría, mientras que age solo requirió una estandarización básica.

Referencias

[1] Moro, S., Cortez, P., & Rita, P. (2014). *Bank Marketing Data Set. UCI Machine Learning Repository.* https://archive.ics.uci.edu/

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: *Machine learning in Python.* Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[3] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors. (2020). SciPy 1.0: *Fundamental algorithms for scientific computing in Python*. Nature Methods, 17(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[4] Waskom, M. L. (2021). seaborn: *statistical data visualization*. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

[5] McKinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference, 51–56. https://doi.org/10.25080/Majora-92bf1922-00a

[6] OpenAI. (2024). *ChatGPT* (versión GPT-5.1). https://chat.openai.com