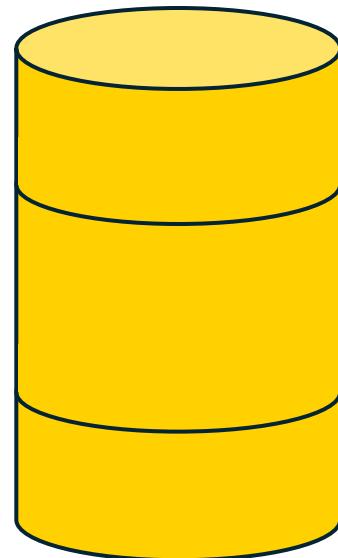


Data Science 2

Profesora: Erica Destefano
Tutor: Federico Gravina
Alumno: Salvador Guagliardi

Bank
Marketing

Uc Irvine
Machine
Learning
Repository



Abstract Bank Marketing

Bank Marketing es un dataset cuyos datos están relacionados a una campaña de marketing directa (llamadas telefónicas) de una institución bancaria portuguesa.

El **objetivo** es predecir si el cliente se suscribirá a un depósito a plazo fijo (variable Y). Era común que se intente contactar más de una vez con el mismo cliente para acceder al producto previamente mencionado (depósito a plazo fijo). Es un análisis **supervisado de clasificación** (se suscribe o no se suscribe) y **bivariado** (solo dos resultados, sí o no).

La variable objetivo, influirá sobre la toma de decisiones de gerentes y gente encargada de diseñar estrategias de retención y fidelización de nuevos clientes. En el análisis, se buscará identificar patrones en los clientes que aceptaron la oferta y, en base a ello, optimizar futuras campañas. Ya que si es una estrategia viable podrán optar por realizarla nuevamente, o descartarla e innovar con algo completamente distinto.

Por lo tanto este proyecto estará **dirigido** para el equipo de marketing de la empresa, data scientists y data analysts, gerentes encargados de la planificación de estrategias comerciales, y gerentes que mantengan relaciones con altos cargos de la empresa.

Link: <https://archive.ics.uci.edu/dataset/222/bank+marketing>



Descripción breve de cada columna

Filas: 45211
Columnas: 17

Unknown son valores desconocidos, podría decirse nulos, que vienen a llenar el vacío que habría en caso de eliminar nulos.

1. Age, edad
2. Job, trabajo u ocupación
3. Marital, es el estado civil: casado, divorciado (incluye viudo/as), soltero
4. Education, es el nivel de educación
5. Default, si tiene crédito en default o no (está atrasado con un pago/debe algo, sí o no)
6. Balance, es el promedio del balance anual
7. Housing, si tiene un préstamo hipotecario sobre su casa
8. Loan, si tiene algún préstamo personal (no hipotecario)
9. Contact, como se le contactó (ejemplo, vía teléfono)
10. Day, último día que se le contactó
11. Month, último mes que se le contactó

Tipos de columnas:

Numéricas: age, balance, day, duration, campaign, pdays, previous.

Categóricas: job, marital, education, default, housing, loan, contact, month, poutcome, Y.

12. Duration, duración del último contacto, en segundos (numerica). Nota importante: este atributo afecta fuertemente el resultado de la variable objetivo (ejemplo, si duration=0 entonces y='no'). Sin embargo, la duración no se sabe antes de realizar la llamada. Así, después de finalizar la llamada recién se sabría si hay "Y". De este modo, esta feature/columna/variable solo debería incluirse para realizar comparaciones y debería de descartarse si la intención es tener un modelo predictivo realista.
13. Campaign, cantidad de veces que se llamó al mismo cliente durante dicha campaña
14. Pdays, cantidad de días pasaron hasta que se le volvió a llamar(si se haya -1 quiere decir que no se le ha contactado previamente)
15. Previous, antes de esta campaña se realizó otra campaña y "previous" contabiliza cuantas veces se le llamó durante dicha campaña anterior
16. Poutcome, (p, previous, outcome, resultado) resultado previo en la campaña anterior, si fue exitoso, falló, o no existió.
17. Y, la variable objetivo (target), va por sí o por no, sí el cliente se suscribió o no se suscribió (probablemente en un futuro deba adaptarse a no = 0, y sí = 1 para machine learning).



Análisis exploratorio inicial (media, desviación estándar, max, min, 25%, 50% y 75%, y para las variables categóricas, conteo y el valor más frecuente).

La columna “Poutcome” tiene 36959 “unknowns” esto quiere decir que hay alrededor de 37 mil personas que se convirtieron en clientes nuevos, ya que no fueron contactados previamente.

La columna “pdays”, tiene -1, en los porcentajes, 25, 50 y 75, como se menciona en el apartado anterior, eso indica, que este es el primer contacto con estos clientes, reforzando el punto de arriba, mostrando que son clientes nuevos.

La edad, al tener una desviación estándar alta (10.6), 18 como valor mínimo y 95 como valor máximo, demuestra que se contactaron clientes de todos los rangos etarios.

De la columna “campaign”, se denota que el 75% de los clientes fueron contactados 3 veces o menos. Y lo más probable es que en el otro 25% se encuentren la mayoría de los outliers, como el máximo, que muestra que se contactó a una persona 63 veces.

La columna “balance”, tiene una desviación estándar altísima, ya que el valor mínimo está 8 mil negativo, y el valor máximo en más de 100 mil. Esto muestra el gran desbalance entre los promedios de balances anuales de los clientes.

La columna, “marital” denota a simple vista que el estatus más común es estar casado.

La variable “education” muestra que la mayoría tiene educación de nivel secundaria, es decir que si se limitaron a aprender lo poco que se aprende de finanzas en el colegio, es probable que no inviertan en algo que desconocen.

La feature “default”, demuestra que la cantidad de gente que potencialmente puede suscribirse y debe dinero es ínfima, casi nula.

El campo “housing”, por otro lado, indica que poco más del 50% se encuentra en un préstamo hipotecario.

Entre los meses, mayo fue el mes más frecuente de todos.

La columna target, “y”, tiene una gran presencia de "no" ya que hay 37k que nunca habían sido contactados, de 39k "no"s, hay solo 2k que si fueron contactados previamente y se siguen negando. Y en promedio, los que aceptaron “yes” (de la target), fueron contactados 2.1 veces contrastando, contra los que no aceptaron “no”, contactados, 2.8 veces. Lo que significa, que contactar más veces a un cliente, aumenta la probabilidad de que rechace la suscripción.



Gráfico de líneas evolutivo

Este gráfico, muestra, la cantidad de veces que se obtuvo un “yes” para la variable objetivo.

Se observa un crecimiento exponencial desde marzo hasta mayo, y recién en septiembre se retorna a un nivel más bajo como lo fue marzo.

La diapositiva anterior, explicaba que mayo fue el mes más con mayor frecuencia de contactos, por lo que sugiere, que esto se puede oponer a la teoría que insistir equivale a ser rechazado, o, que la empresa, durante el primer trimestre del año, tuvo una fase más tranquila y planificadora. Y por último, no descartar que algún factor externo pudo haber influido en dicho resultado.

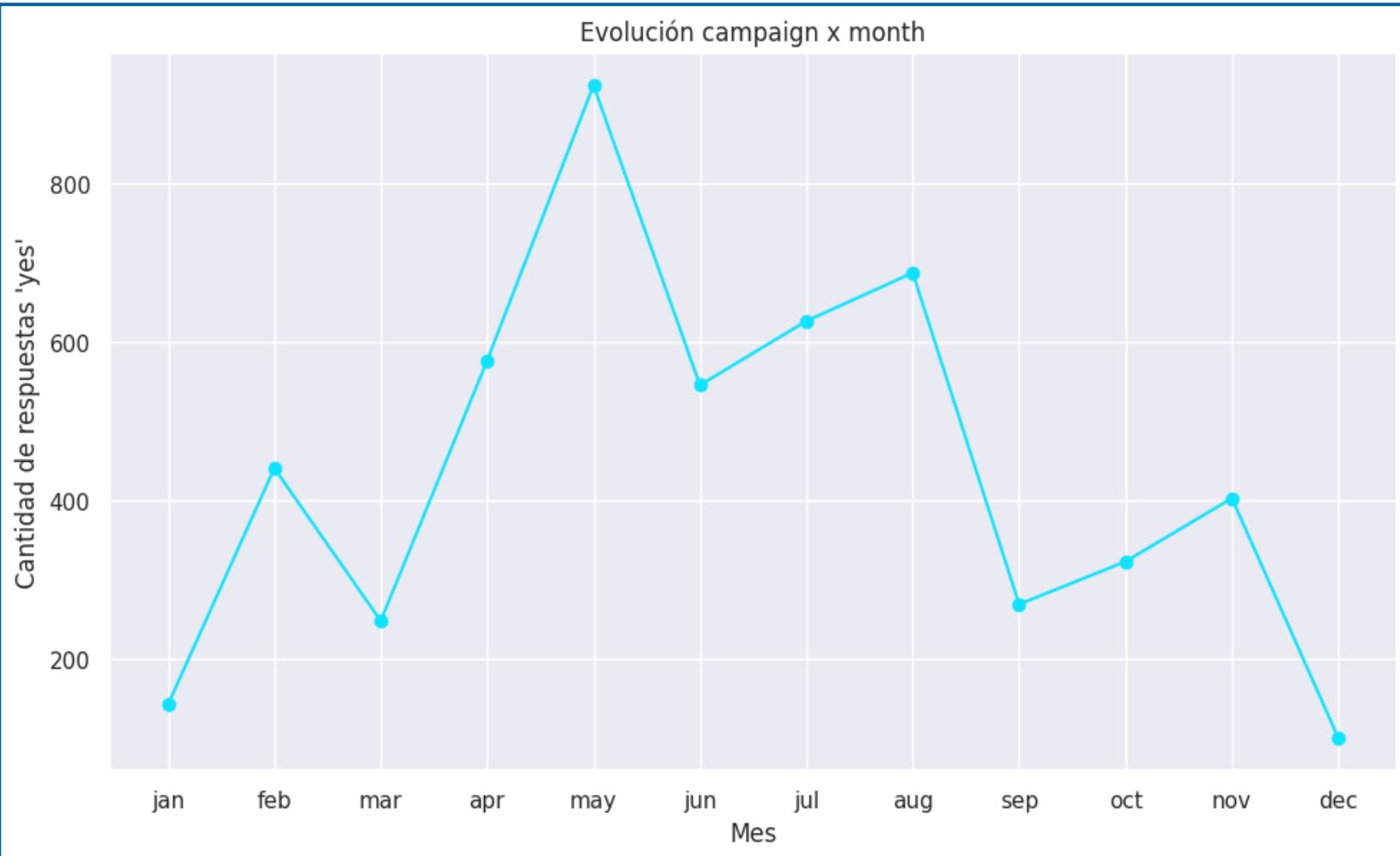




Gráfico de boxplot

En este boxplot, se observan claramente la gran cantidad de outliers, ese otro 25% que fueron contactados más de 3 veces. Es lógico deducir que a cualquier persona le molestaría tener más de 10 llamadas de un banco para venderte un servicio.

Esto demuestra la ineficiencia de los llamados por parte de la entidad bancaria. En vez de llamar repetidas veces a un mismo cliente, podrían haber invertido ese esfuerzo, en captar nuevos clientes.

Aparte de esto, la gran cantidad de outliers, afecta drásticamente el valor promedio, dificultando la posibilidad de obtener información a través de dichos datos.

Sin embargo, mirando el lado positivo, el otro gran 75% de los clientes fueron contactados 3 veces o menos. Manteniendo así un margen más respetable de llamadas a una misma persona.

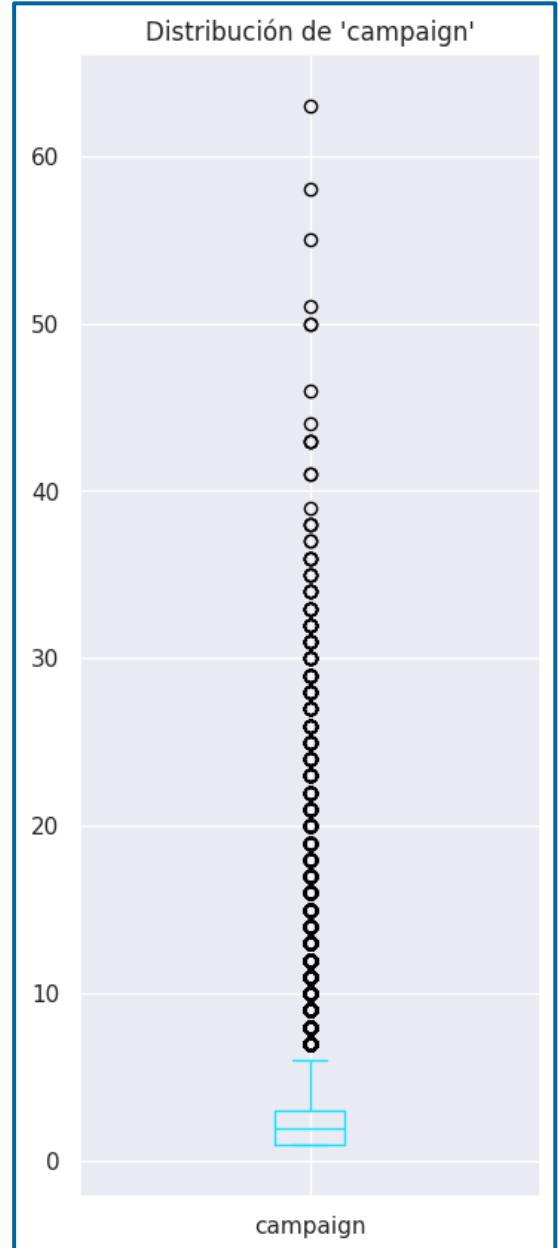


Gráfico de barras

Este gráfico de barras, muestra que los estudiantes, jubilados y desempleados, son las personas con en el menor promedio de llamadas, ya que es probable que mantengan un balance más bajo que las demás ocupaciones.

Por otro lado, “unknown” muestra que la gran mayoría de personas contactadas tienen una ocupación que no encaja con las etiquetas seleccionadas, o no quisieron revelar su ocupación porque sintieron que era muy invasivo contar eso.

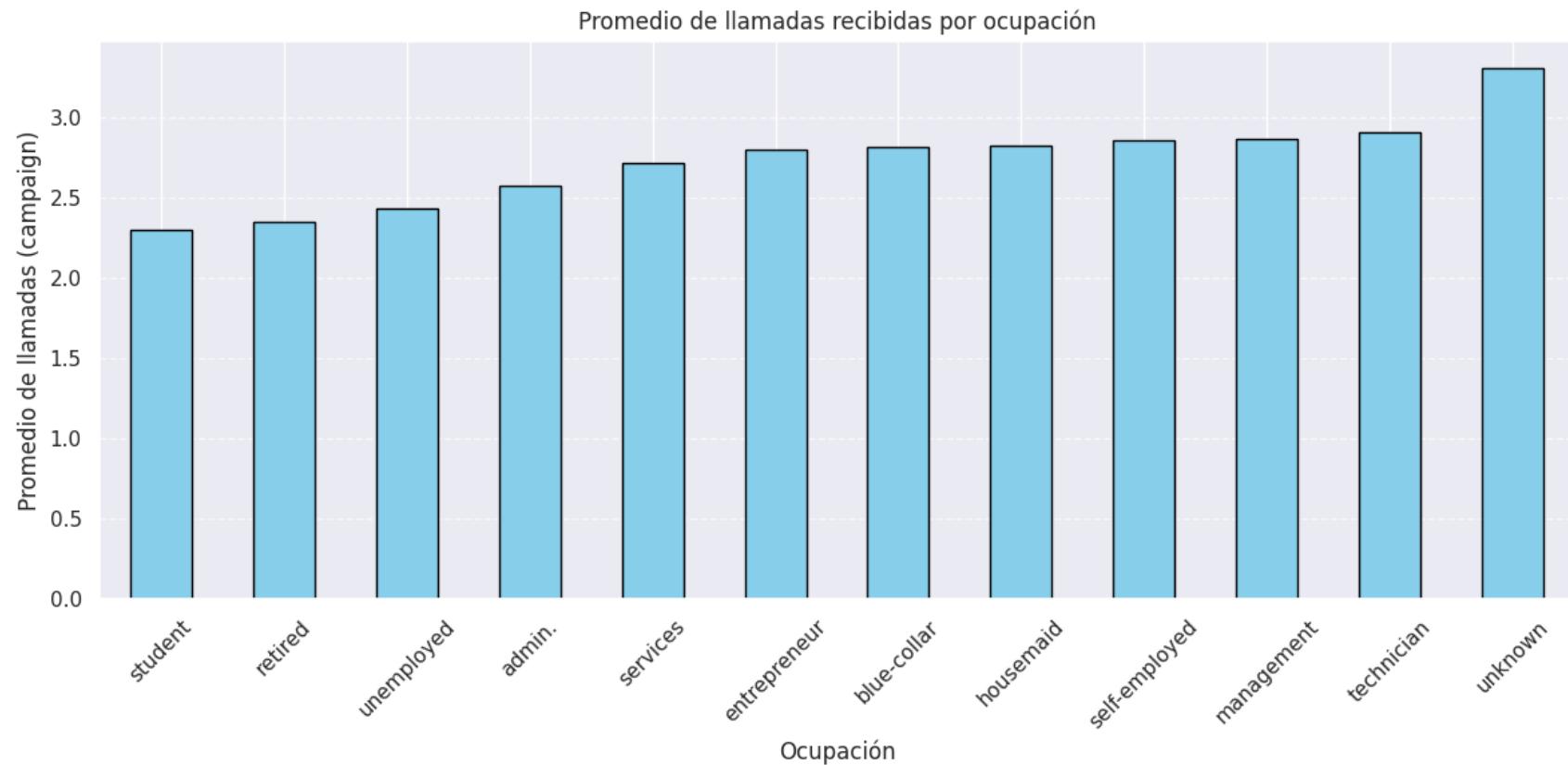




Gráfico de dispersión

Este gráfico de dispersión, denota que las edades entre 40 y 50 son las más presentes en el gráfico. Por otro lado, el rango entre 50 y 60, son los que parecen tener mayor balance, y sin embargo, al tener más balance que los demás, siguen rechazando la oferta. Indicando que por más que tengan dinero, si la oferta no les resulta atractiva, no van a invertir y pagar el servicio de suscripción.

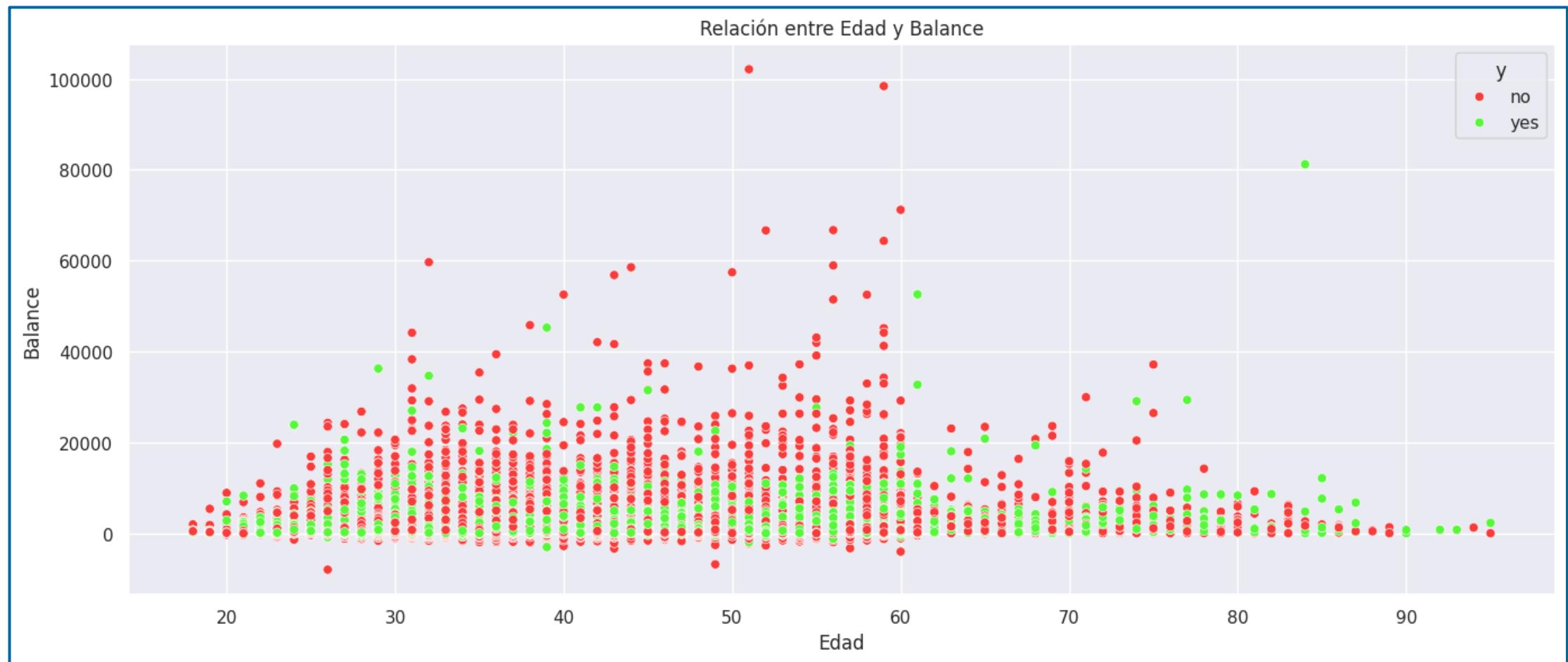




Gráfico de barras

Este gráfico de barras muestra una transición gradual entre los estados civiles. El estado “married” o casado, muestra que son los más propensos a decir que no, tanto como sí. Ya que al estar en pareja y casado, se asume que ambas personas tienen responsabilidad financiera y pueden pagar el servicio ofrecido. Sin embargo, el estado “single” o soltero, muestra que puede no tener la mayor cantidad de “yes” pero sí la mayor cantidad de éxito, ya que la barra verde, es casi igual de alta a la de los casados, y la barra roja (los “no”) de los casados es mucho más alta que la de los solteros. Todo esto se traduce en que al venderle el servicio a soltero/a, hay una mayor probabilidad de éxito que contra alguien casado.

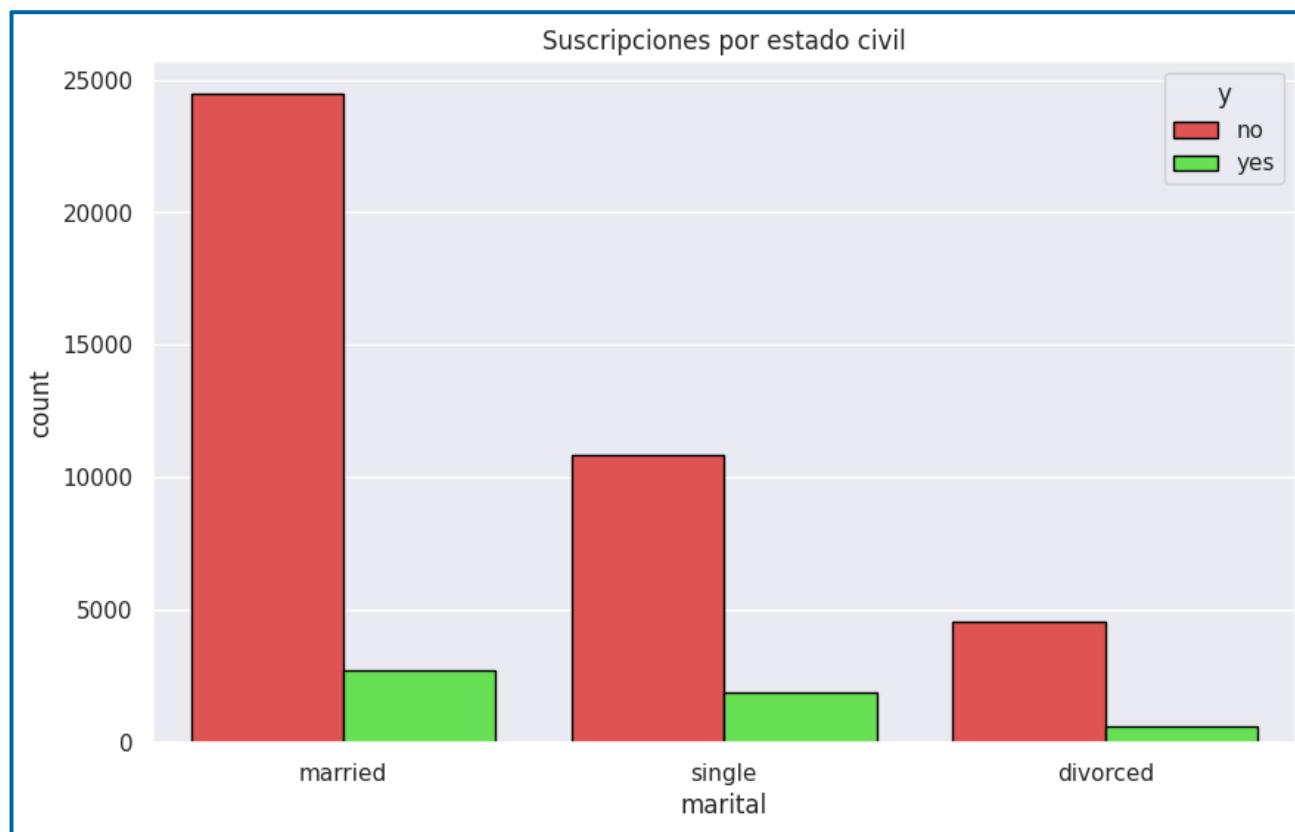
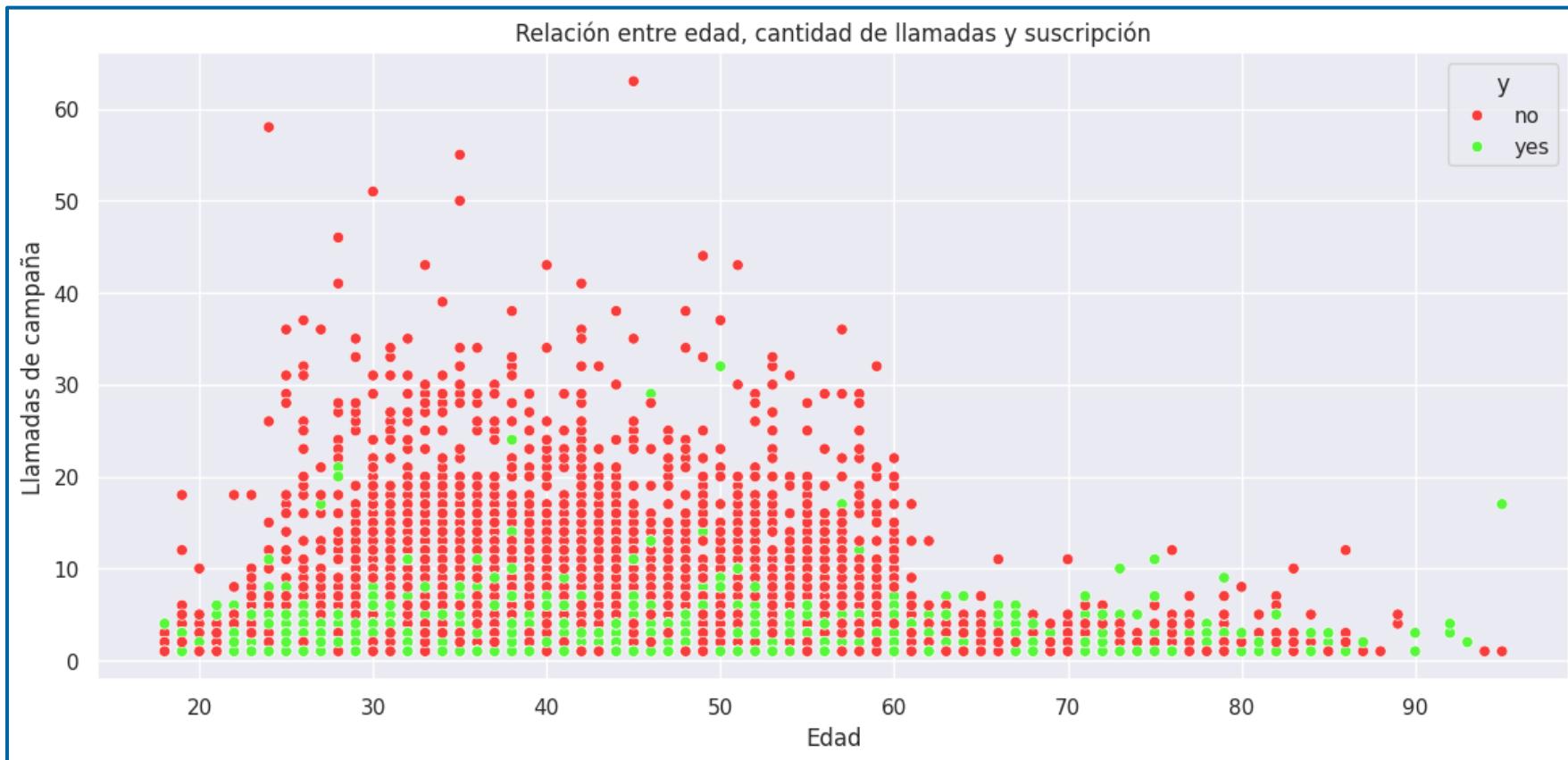




Gráfico de dispersión

Este gráfico de dispersión, se ve que en el eje x, las edades de 30 a 40 es la que mayor cantidad de burbujas tiene (tanto normales como outliers), mostrando que ahí se ubica la mayoría de la clientela. Con esto en mente, no hay que ignorar, que las edades de 20 a 30 parecen tener una mayor proporción de burbujas verdes contra rojas, mostrando que son los clientes más propensos a pagar la suscripción. Trasladándonos al eje Y, se ve que la gran mayoría de burbujas verdes, se encuentran en el rango 0-10, sin embargo, hay excepciones inclusive una por arriba de la línea del 30. Esto significa que las personas son más propensas a pagar el servicio cuando se les insiste menos.





Conclusiones obtenidas de los gráficos y análisis exploratorios iniciales

La institución bancaria podría optar por apuntar más a las personas, solteras, en un rango de edad de 20 a 30 años y con un balance, aceptable que pueda permitirse pagar el servicio. Sin insistir demasiado, con 2 intentos de llamadas telefónicas, podría ser suficiente y una tercera para cerrar el trato.

En cuanto a lo negativo, la empresa definitivamente podría replantearse su esquema de llamados, ya que es totalmente ilógico que haya un 25% de la clientela que haya sido contactada más de 3 veces, y unos cuantos pares de clientes, superando los 20 contactos. Es demasiado, y en todo caso, esos clientes son irrecuperables.

No hay que descartar la posibilidad, de que hay una gran cantidad de clientes nuevos, que todavía tengan que desarrollar, cierto lazo, con la empresa y poder confiarle lo suficiente como para pagar la suscripción. Se podría invertir en otra campaña de para formar clientes más leales, como alternativa, ya que clientes más leales a la larga se traducen en potenciales compradores del servicio ofrecido.

Se recomienda abrir el Google Colab para poder comparar y sacar conclusiones propias, si desea realizar un análisis más profundo y cuestionar mis insights.

https://colab.research.google.com/drive/1P9j37y-luasl5iATOr113uH_o7zHRoq-?usp=sharing



Modelado

Los primeros 3 modelos fueron entrenados sin ningún tipo de optimización, ni hiper parámetros y en base a estos 3 modelos “en crudo” se pudieron obtener conclusiones y datos que ayudarían a más adelante realizar el modelado optimizado.

Los 3 modelos mencionados son el árbol de decisión, KNN (“K nearest neighbor”) y la regresión logística, y a continuación se listarán las conclusiones obtenidas en base a las métricas “accuracy” (exactitud), “precision” (precisión), “recall” (sensibilidad/exhaustividad), “F1 Score”, “Train Time” (tiempo de entrenamiento) y ROC-AUC(Receiver Operating Characteristic-Area Under the Curve):

Accuracy: Mide la proporción de predicciones correctas con respecto al total de predicciones, pero al tener un desbalance muy alto entre clases, no predice correctamente la clase minoritaria, que es el “Sí”, a pesar de ser “la mejor” métrica.

Precision: Mide cuán precisas son las predicciones positivas del modelo, es decir, de todas las instancias que el modelo ha predicho como positivas, cuántas de ellas realmente lo son. Es bastante engañosa, porque los positivos falsos, son 3 veces la cantidad de los positivos verdaderos. Y si predice que es positivo, pero en realidad no lo es, no termina siendo muy útil. Salvo en el caso de KNN, que está dentro del todo balanceado.

Recall: El recall mide cuántos de los casos que son realmente positivos fueron identificados como tales por el modelo. Considero que esta métrica es la que mejor indica que tan acertados son estos modelos. Ya que prioriza la clase minoritaria (el Sí), destacando el rendimiento de la regresión logística, aunque sigue sin ser una métrica muy positiva y alentadora.

F1_score: Combina la precisión y el recall en un único valor, proporcionando un equilibrio entre ambas. Esta métrica, prueba mi punto anterior, que el modelo no es el mejor, un promedio aproximado entre los 3 modelos, debe ser de 0.33, y como se sabe, mientras más cercano sea el valor a 1 significa que mejor es el equilibrio entre precisión y recall. Y 0.33 tiende más para el lado de cero, no llega a ser tan pésimo como un 0.20 de todas formas.

Train_time: El modelo más rápido fue regresión logística, y el más lento, KNN.

Roc_auc: El mejor modelo es regresión logística. Teniendo en cuenta que el 0.5 es azar, y el 1 perfección (cosa que es imposible, ya que si uno tira una moneda al aire, nunca acierta con un 100% de precisión de que lado va a caer), la regresión logística, está dentro del todo bien, en el punto medio, y los demás modelos implican que se deben de hacer ajustes.

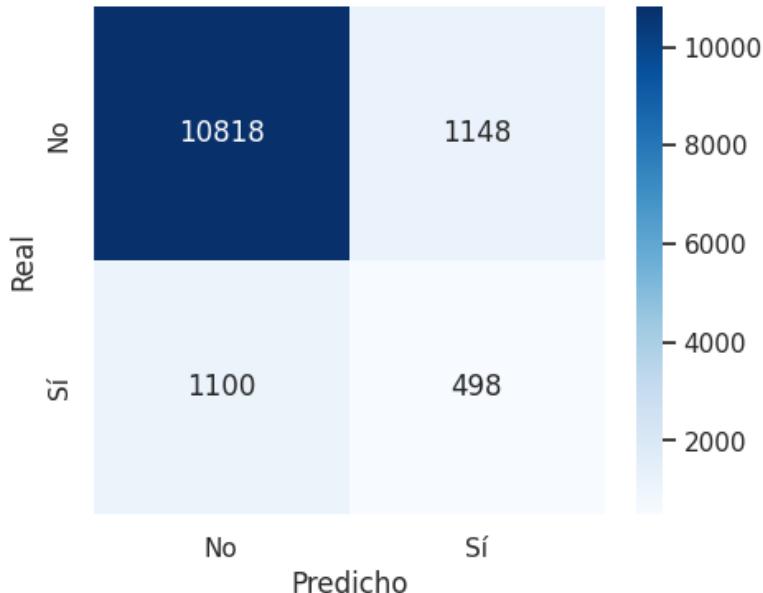


Sugerencias

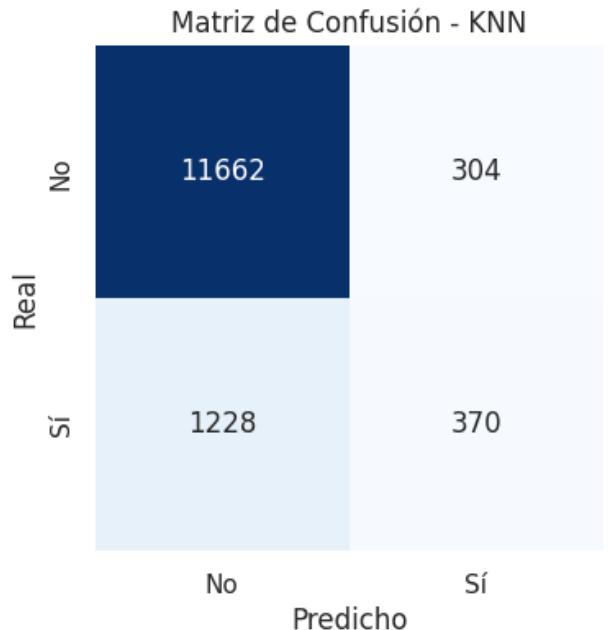
En base al análisis previo, a continuación se deja una breve sugerencia:

Recaudar más datos para tener una mayor cantidad de datos de la clase minoritaria y así poder encontrar más patrones que ayuden al modelo a ser entrenado y progresar. Esto como efecto secundario tiene reducir el gran desbalance de clases que hay. Y hablando de overfitting/sobre ajuste, que está ligado al desbalance mencionado anteriormente, realizar una validación cruzada, podría ayudar a ajustar el sobre ajuste y que el modelo realmente sirva. Y por último, recomendaría utilizar algún modelo similar a la regresión logística, o modelos populares como xgboost o random forest classifier o lightGBM.

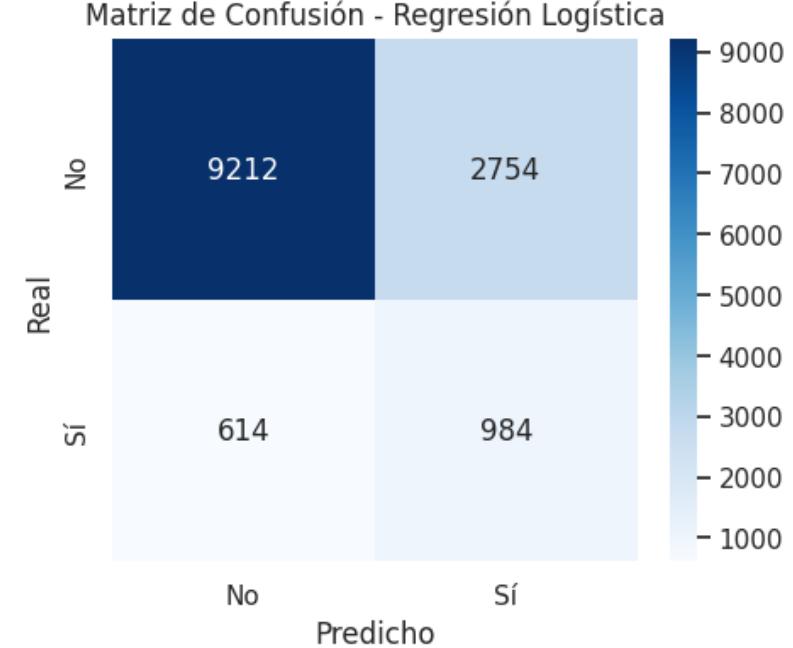
Matriz de Confusión - Árbol de Decisión



Matriz de Confusión - KNN



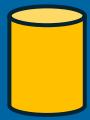
Matriz de Confusión - Regresión Logística



Resumen de métricas

	accuracy	precision	recall	f1	train_time
Árbol de Decisión	0.834267	0.302552	0.311640	0.307028	0.396001
KNN	0.887054	0.548961	0.231539	0.325704	0.020084
Regresión Logística	0.751696	0.263242	0.615770	0.368816	0.526007

	roc_auc
Árbol de Decisión	0.607851
KNN	0.699110
Regresión Logística	0.761576



Modelado optimizado

A continuación, se decidieron aplicar 3 modelos distintos a los previamente aplicados con las optimizaciones, y validaciones cruzadas. Los 3 modelos "antiguos" seguirán formando parte del análisis ya que proveen buenos insights y son los que dieron el hincapié a los 3 nuevos modelos optimizados:

Regresión logística + grid search CV

Random forest + randomized search CV

XGBoost + randomized search CV

Estos 3 modelos fueron seleccionados en torno a intentar obtener más detalles en lo que respecta al "Sí" en la variable objetivo (la clase minoritaria) que al tener un dataset desbalanceado, es lo más complicado de encontrar y predecir de manera correcta, contrario al caso del "No". Y las métricas utilizadas para evaluar su rendimiento serán: "precisión", "recall", "F1-Score", "macro average" (promedio macro) y ROC-AUC. Nota: en el notebook de Colab se pueden observar "accuracy" y "weighted average" pero debido a que "maquillan" la realidad, se obviaron para el siguiente análisis. "Weighted average" o promedio ponderado, calcula un promedio en torno a la clase más grande y dominante, a diferencia del "macro avg", que muestra la realidad "cruda", y por otro lado, "accuracy" como fue mencionado previamente, no predice correctamente la clase minoritaria.

ROC-AUC: XGBOOST con un ROC-AUC de 0.7970648533381084, es el modelo con mejor performance de los 3, los otros 2 están muy cerca y gracias a la optimización y demás alcanzaron un resultado tan notable. (Regresión Logística (optimizada): 0.76104009336424 y Random Forest: 0.7921073621820021). Recordar que el ROC-AUC mide qué tan bien un modelo separa a los que van a decir "Sí" de los que van a decir "NO", o sea, muestra la capacidad real de discriminación del modelo. (Nota menor: regresión logística aumentó su rendimiento un 0.03 gracias a la optimización).

El "macro average" muestra un rendimiento moderado y/o balanceado del modelo, evidenciando que su capacidad para clasificar ambas clases de forma equilibrada es limitada, principalmente por el bajo "recall" en la clase positiva (el "recall" en la regresión logística, la clase 0 tiene 0.99, y la clase 1 tiene 0.18, eso tira muy para abajo el promedio). Por otro lado, el "weighted average" muestra valores elevados (en relación al macro avg) debido al fuerte predominio de la clase negativa, lo que genera una percepción optimista del modelo que no refleja fielmente su capacidad para detectar clientes potenciales (tomando el ejemplo de la regresión logística, el macro avg es de 0.59 y el weighted avg es de 0.89).



Modelado optimizado

(“Support” = cantidad real de ejemplos de cada clase en el conjunto de test)

Como se observa en las imágenes, las métricas previamente explicadas en los modelos no optimizados, como “precisión”, “recall”, “F1-Score”, son todas muy similares en los 3 modelos de Machine Learning aplicados, y es por eso que se le dio más importancia al ROC-AUC en la diapositiva anterior, ya que esa métrica es la que precisamente muestra el modelo que mejor rinde (es decir, el modelo que tiene mejor performance).

Lo que cabe destacar, es que estos 3 modelos aplicados con optimizaciones incluidas, tienen un rendimiento superior a los no optimizados (Árbol de decisión 0.60 de ROC-AUC y KNN 0.69 de ROC-AUC).

Mirando el “F1-Score” de la clase 1, (recordar que este número, mientras más cercano a 1 mejor equilibrio hay entre precisión y recall) ya alcanza para determinar que XGBoost es el que mejor equilibrio tiene a diferencia de los demás.

Llegando a la conclusión final, si hay que elegir un modelo, sin dudas cabe optar por XGBoost y seguido de este, Random Forest. Y las sugerencias no difieren de las mencionadas anteriormente, sin dudas el gran desbalance entre clases es un factor limitante y sería crucial recaudar mayores volúmenes de datos para así en análisis futuros poder obtener conclusiones más precisas sobre la clase minoritaria que es la clase objetivo y lograr comprender que hace que el cliente de el “sí”.

REGRESIÓN LOGÍSTICA OPTIMIZADA

	precision	recall	f1-score	support
0	0.90	0.99	0.94	11966
1	0.66	0.18	0.29	1598
accuracy			0.89	13564
macro avg	0.78	0.59	0.61	13564
weighted avg	0.87	0.89	0.86	13564

RANDOM FOREST OPTIMIZADO

	precision	recall	f1-score	support
0	0.90	0.99	0.94	11966
1	0.68	0.21	0.32	1598
accuracy			0.90	13564
macro avg	0.79	0.60	0.63	13564
weighted avg	0.88	0.90	0.87	13564

XGBOOST OPTIMIZADO

Reporte:

	precision	recall	f1-score	support
0	0.90	0.98	0.94	11966
1	0.64	0.23	0.33	1598
accuracy			0.89	13564
macro avg	0.77	0.60	0.64	13564
weighted avg	0.87	0.89	0.87	13564

ROC AUC REGRESIÓN LOGÍSTICA: 0.76104009336424 #3

ROC AUC RANDOM FOREST: 0.7921073621820021 #2

ROC AUC XGBOOST: 0.7970648533381084 #1

Bank Marketing

Muchísimas gracias por tomarte el tiempo de mirar y leer esta presentación con toda la documentación del Colab, sumado a mirar el notebook, espero que te haya gustado mi análisis!

Este es de mis primeros análisis como data scientist, aplicando líneas de código, limpieza de datos, encoding, gráficos, etc, todo lo visto anteriormente.

Todas las críticas son bienvenidas, siempre me sirve obtener una perspectiva distinta sobre como encarar el dataset, ya sea una línea de código, un análisis, o inclusive conclusiones que se me pueden haber pasado por alto mientras me enfocaba en otras.

De nuevo, gracias por tu tiempo, y espero que te haya gustado y si te sirve de inspiración, aún mejor!

Link Colab notebook: https://colab.research.google.com/drive/1P9j37y-luasl5iATOr113uH_o7zHRoq?usp=sharing

Link Git Hub: <https://github.com/salvadorGdi/Data-Science-2>

Link del dataset: <https://archive.ics.uci.edu/dataset/222/bank+marketing>