ESTUDIO DE MERCADO: BANCO DE PORTUGAL

Trabajo final – Data Science (CoderHouse)

Descripción breve

Análisis de Base de Datos y Entrenamiento de Modelos de Machine Learning como para resolver problemáticas económicas.

Victoria Gardella Ruiz, Ornella Padini, Pablo Rabal
v.gardella.ruiz@gmail.com

Contenido

Descripción del caso: 3

Nuestro cliente: ¿Cómo podemos ayudarlo? 3

Objetivos de la investigación: 3

Descripción de los datos: 4

Descripción General: 4

Atributos: 4

Manipulación y limpieza de los registros: 5

Exploratory Data Analysis (EDA). 7

Análisis Univariado: Análisis de la Población. 7

- Variable objetivo: Consolidación de un plazo fijo. 7
- Puesto de trabajo. 8
- Educación. 8
- Edad. 9
- Estado civil. 9

Análisis Univariado: Interacción con los clientes 10

- Número y duración de los contactos con los clientes. 10
- Contactos telefónicos por mes. 11

Análisis Bivariado. 12

- Préstamos Personales e Hipotecarios. 12

....12

Análisis Multivariado. 14

- Edad de los clientes, Duración del último contacto y Constitución de Plazo Fijo. 14
- Tenencia de Préstamos y Constitución de Plazo Fijo. 15
- Circunstancias del último contacto y constitución de plazos fijos. 16
- Edad de los clientes, último contacto y constitución de un Plazo Fijo. 18

Selección de las variables. ....**¡Error! Marcador no definido.**

Entrenamiento y selección del modelo de clasificación. 19

- Selección de los modelos. 19
- PyCaret. 19
- Scikit-Learn 21

Optimización del modelo 23

- Selección de variables. 23
- Ajuste de Hiperparámetros 27

Resultado Final y Conclusiones.....28

Descripción del caso:

Nuestro cliente: ¿Cómo podemos ayudarlo?

Nuestro cliente, el Banco de Portugal, ha encontrado una disminución en sus ingresos. Luego de investigar al respecto se descubrió que el problema radicaba en que los clientes no estaban invirtiendo lo suficiente en plazos fijos a largo plazo. Las autoridades han decidido identificar cuáles de los clientes existentes tienen una mayor posibilidad de constituir dichos depósitos, de forma de enfocar los esfuerzos de marketing en ellos.

Se nos ha encomendado la tarea de analizar el perfil de los clientes de dicho banco; buscaremos datos que nos puedan servir para generar estrategias de marketing y generaremos modelos de clasificación de los clientes a fin de permitirle al Banco de Portugal el enfocar mejor sus recursos y obtener los mejores resultados económicos.

Objetivos de la investigación:

Se desea determinar qué clientes de este banco son posibles suscriptores a un depósito a largo plazo, también conocido como Plazo Fijo. Para realizar estas predicciones analizaremos a que otros servicios financieros se encuentran suscriptos y algunas de sus características personales.

Descripción de los datos:

Descripción General:

Los datos de los que disponemos provienen de campañas de marketing directo basadas en llamadas telefónicas. En muchos casos se necesitó más de un contacto con el cliente para determinar si este se suscribiría a un depósito a largo plazo o no.

La base de datos se encuentra en formato "csv" y cuenta con dos conjuntos de datos, uno de entrenamiento de aproximadamente 33.000 registros y 16 atributos, y uno de testeo de aproximadamente 8.000 registros y 15 atributros, el cual no será usado para hacer el análisis. La base de datos de entrenamiento contiene un atributo único denominado "Y", el cual indica si el resultado deseado (la suscripción del cliente a un depósito a largo plazo o plazo fijo) es positivo ("yes") o negativo ("no").

Atributos:

El conjunto de trabajo, como mencionamos antes, cuenta con un total de 16 atributos: 15 variables que nos brindaran información sobre nuestros clientes y una variable que usaremos como blanco de nuestra investigación.

1. Age (numérico) - Edad.
2. Job (categórico, nominal) - Trabajo.
3. Marital (categórico, nominal) - Estado civil.
4. Education (categórico, nominal) - Educación.
5. Default (categórico, nominal) - Crédito en default.
6. Housing (categórico, nominal) – ¿Tiene un préstamo hipotecario?.
7. Loan (categórico, nominal) – ¿Tiene préstamos personales?.
8. Contact (categórico, nominal) – Forma de contacto (teléfono o móvil).
9. Month (categórico, nominal) – Ultimo mes de contacto.
10. Dayofweek (categórico, nominal) – Ultimo día de contacto.
11. Duration (numérico) – Duración del último contacto en segundos.
12. Campaign (numérico) – Número de contactos durante una campaña.
13. Pdays (numérico) – Número de días desde el último contacto.
14. Previous (numérico) – Número de contactos durante la campaña anterior.
15. Poutcome (categórico, nominal) – Rédito de la última campaña (¿Fue exitosa?).

Target:

1. Y (binario) – ¿Se suscribió el cliente a un préstamo a largo plazo?

Manipulación y limpieza de los registros:

La base de datos con la que trabajaremos fue analizada y limpiada por la entidad otorgante con anterioridad; esto quiere decir que no contamos con registros con valores nulos. Se analizó la presencia de valores duplicados, obteniéndose un total de 8 posibles registros; sin embargo, ante la falta de alguna variable descriptiva o clave identificatoria (un ID de cliente, por ejemplo), decidimos no eliminar dichas entradas.

Se editó la base de datos de forma de transformar las variables objeto en variables de tipo categórico y se armo un pequeño diccionario con las claves de las cuales disponíamos:

Variable	Claves
'Job' (Trabajo)	0: 'admin.', 1: 'blue-collar', 2: 'entrepreneur', 3: 'housemaid', 4: 'management', 5: 'retired', 6: 'self-employed', 7: 'services', 8: 'student', 9: 'technician', 10: 'unemployed', 11: 'unknown'
'Marital' (Estado Civil)	0: 'divorced', 1: 'married', 2: 'single', 3: 'unknown'
'Education' (Nivel Educativo)	0: 'basic.4y', 1: 'basic.6y', 2: 'basic.9y', 3: 'high.school', 4: 'illiterate', 5: 'professional.course', 6: 'university.degree', 7: 'unknown'
'Default' (¿Tiene Credito en Default?)	0: 'no', 1: 'unknown', 2: 'yes'
'Housing' (¿Tiene crédito Hipotecario?)	0: 'no', 1: 'unknown', 2: 'yes'
'Loan' (¿Tiene Prestamos Personales?)	0: 'no', 1: 'unknown', 2: 'yes'
'Contact' (Método de Contacto)	0: 'cellular', 1: 'telephone'
'Month' (Mes del último contacto)	0: 'apr', 1: 'aug', 2: 'dec',

	3: 'jul', 4: 'jun', 5: 'mar', 6: 'may', 7: 'nov', 8: 'oct', 9: 'sep'
'Day_Of_Week' (Día del último contacto)	0: 'fri', 1: 'mon', 2: 'thu', 3: 'tue', 4: 'wed'
'Poutcome' (Resultado de la última Campaña)	0: 'failure', 1: 'nonexistent', 2: 'success'
'Y' (Variable Target: ¿Se definió un plazo fijo?)	0: 'no', 1: 'yes'

Tabla 1: Diccionario de Claves. Al realizar la codificación de las variables categóricas de las que disponemos se generó un diccionario que contuviera cada una de las variables y las claves asignadas a cada uno de los posibles valores.

Inicialmente se realizó un análisis utilizando la totalidad de las columnas. Posteriormente se utilizaron herramientas de selección de variables (bibliotecas BorutaPy y Boruta Shap) para determinar si los resultados obtenidos al entrenar e implementar los algoritmos de clasificación podían ser refinados.

Análisis Exploratorio de Datos (EDA).

Vamos a iniciar el análisis de la base de datos armando un perfil de los clientes del banco (análisis univariado). El primer paso será analizar cuantas personas constituyeron un plazo fijo, siendo que esta es la variable blanco de nuestro trabajo. Posteriormente analizaremos las características de la población sobre la que vamos a trabajar (distribuciones de los clientes con respecto a su educación, puesto de trabajo, estado civil y edad) y los patrones de interacción entre los empleados del banco y sus clientes durante la última campaña (número, duración y distribución temporal de las llamadas realizadas por los empleados).

Posteriormente vamos a intentar determinar las relaciones entre las características previamente analizadas (análisis bivariado y multivariado) tratando de encontrar patrones que puedan ser de utilidad para nuestro cliente.

Análisis Univariado: Análisis de la Población.

- Variable objetivo: Consolidación de un plazo fijo.

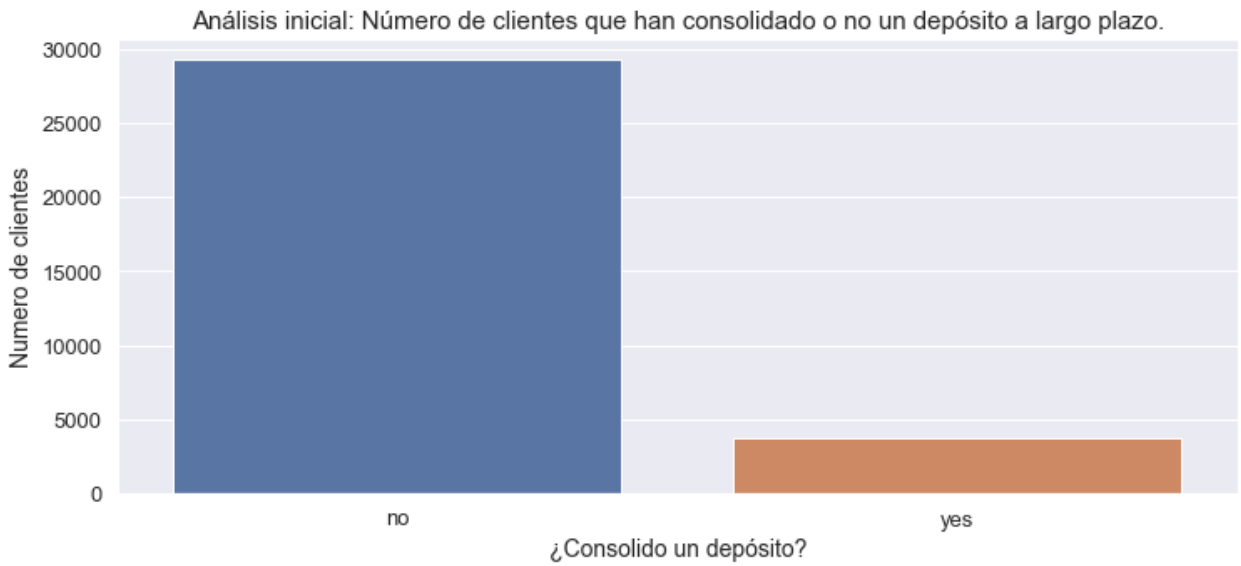

Figura 1: Variable objetivo. Se determina la proporción de clientes que han generado plazos fijos.

Al analizar la distribución de registros de la variable objetivo se observó que el 89% de los clientes del banco (un total de 29238 individuos) generaron depósitos a largo plazo, mientras que el 11% (3712 individuos) no lo hicieron. Debido a esta distribución tan dispareja se deberán tomar ciertas medidas precautorias para evitar obtener errores al entrenar los algoritmos de clasificación.

- Puesto de trabajo.

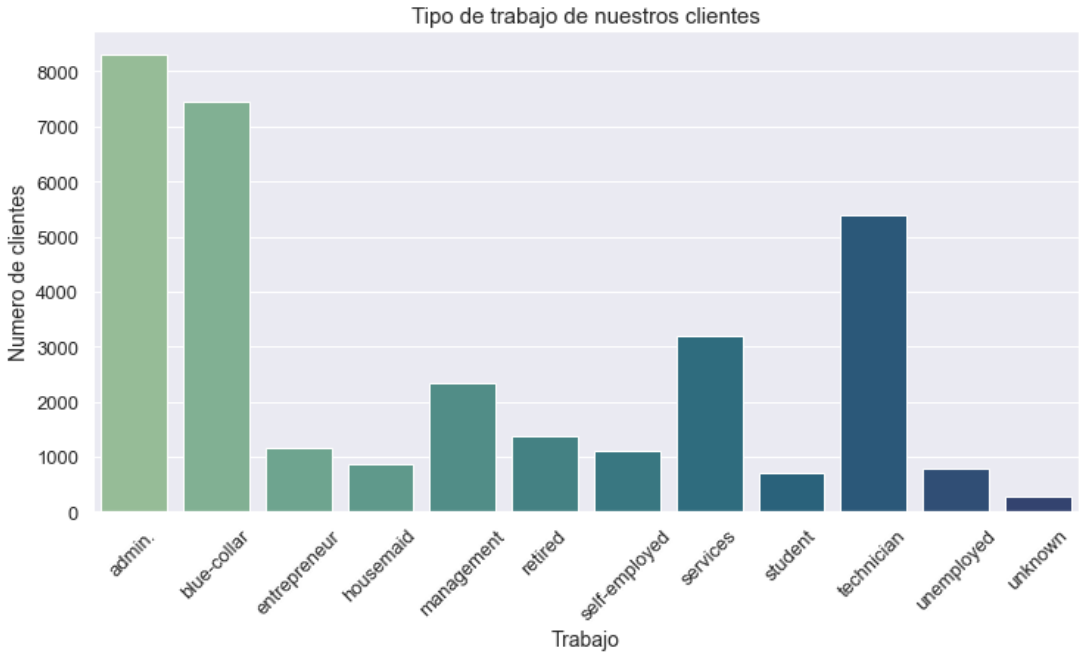

Figura 2: Puestos de Trabajo. Definimos cuales son los tipos de trabajo prevalentes entre la población.

Se puede observar que la mayoría de los clientes del banco trabajan en relación de dependencia, dividiéndose mayormente entre trabajos administrativos (25,2%), de cuello azul (obreros, trabajo manual; 22,6%) y trabajos técnicos (16,4%). Una pequeña proporción corresponde a estudiantes, personas retiradas o desocupados (correspondiente a un total del 8,7%).

- Educación.

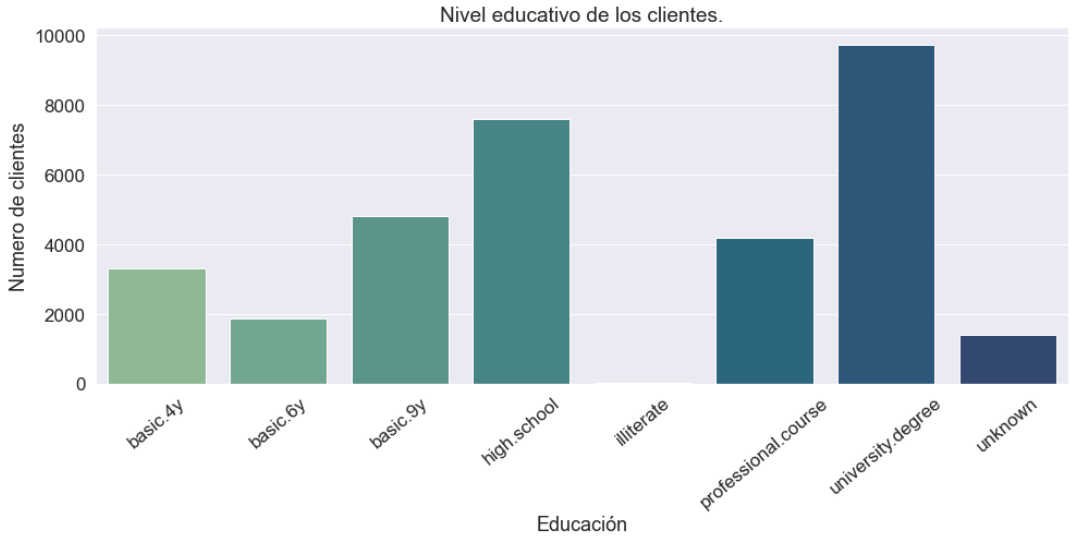

Figura 3: Nivel educativo. Distribución del nivel educativo de los clientes del banco.

El 65,3% de los clientes presentan educación secundaria o superior (29,6% con título universitario, 23% con título secundario y un 12,7% con cursos profesionales realizados).

- Edad.

Al analizar la distribución de edades de los clientes se observa que la mayoría de la población corresponde a personas de entre 20 y 40 años, seguido por una cantidad menor pero todavía sustancial correspondiente a las personas entre 40 y 60 años. Por otro lado, la población correspondiente a personas de la tercera edad (60 años o más) representa una marcada minoría.

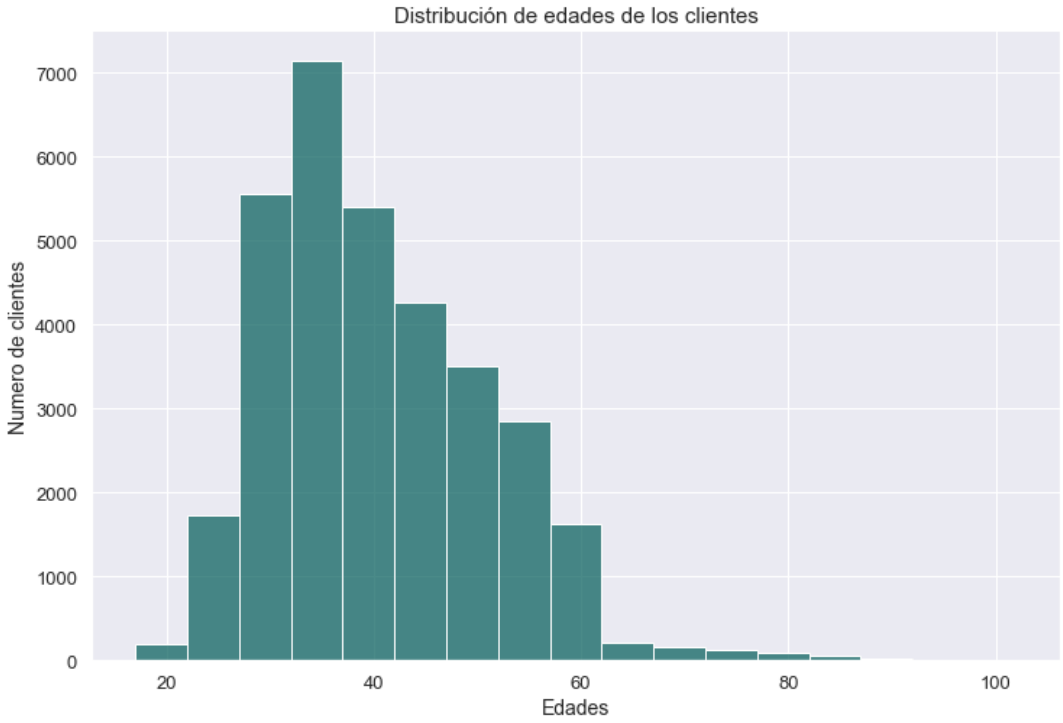

Figura 4: Distribucion de edades.

- Estado civil.

Con respecto al estado civil, el 60,6% de la población se encuentra casada, seguida de un 28,1% correspondiente a personas solteras y un 11,2% correspondiente a personas divorciadas o viudas.

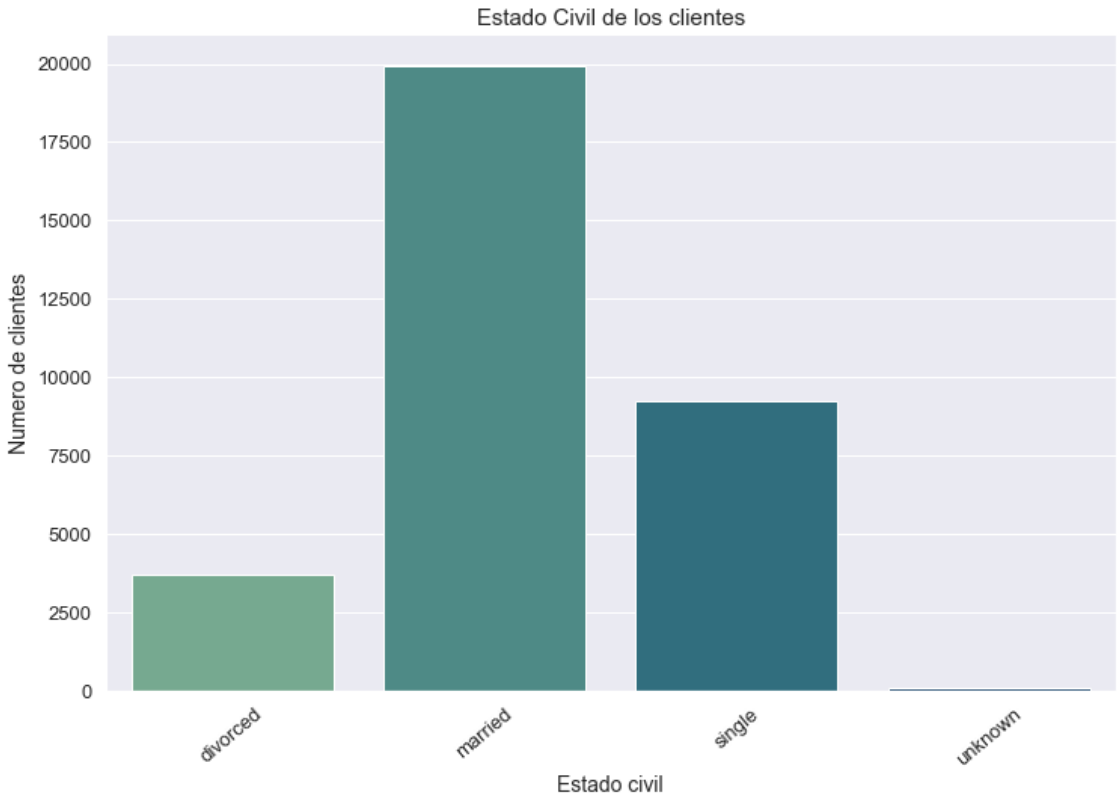

Figura 5: Distribución con respecto al Estado civil.

Análisis Univariado: Interacción con los clientes.

- Número y duración de los contactos con los clientes.

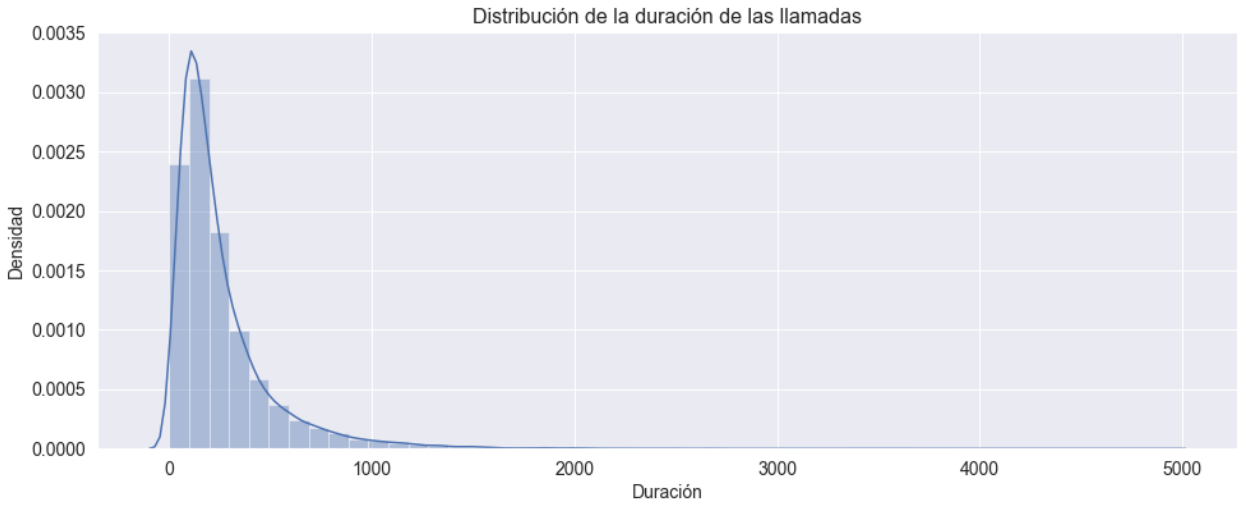

Figura 6: Distribución de la duración de los contactos (llamadas telefónicas).

Cuando graficamos la distribución correspondiente a la duración de los contactos (llamadas entre empleados del banco y clientes, medidas en segundos) se observa que la gran mayoría corresponde a llamadas cortas, desde unos pocos segundos hasta alrededor de 15 minutos. Sin embargo, aunque es difícil de observar, también encontramos algunas llamadas de larga y muy larga duración.

- Contactos telefónicos por mes.

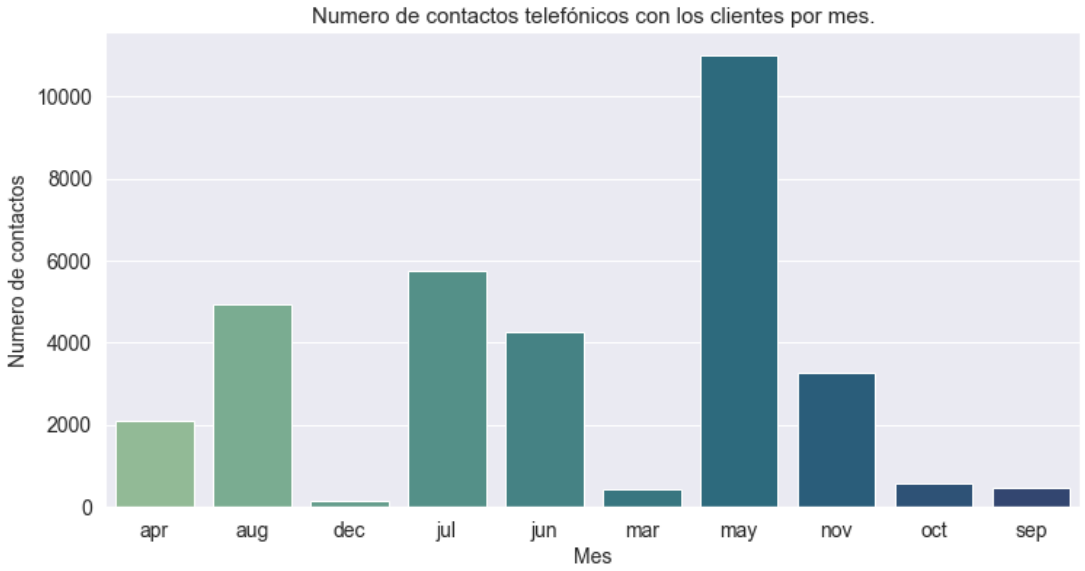

Figura 7: Número de contactos con respecto al mes.

Al cuantificar en que mes se realizó el último contacto para cada uno de los clientes de nuestra base de datos se observa una marcada mayoría correspondiente al mes de mayo, seguido por los meses de julio y agosto. Si analizamos la distribución con respecto al año, se puede ver que desde marzo hasta mayo hay un considerable aumento de las llamadas, disminuyendo en el mes de junio y manteniéndose constante hasta el mes de agosto. Exceptuando el mes de noviembre, se puede ver que el número de contactos disminuye de forma sustancial a partir de septiembre.

Análisis Bivariado.

- Préstamos Personales e Hipotecarios.

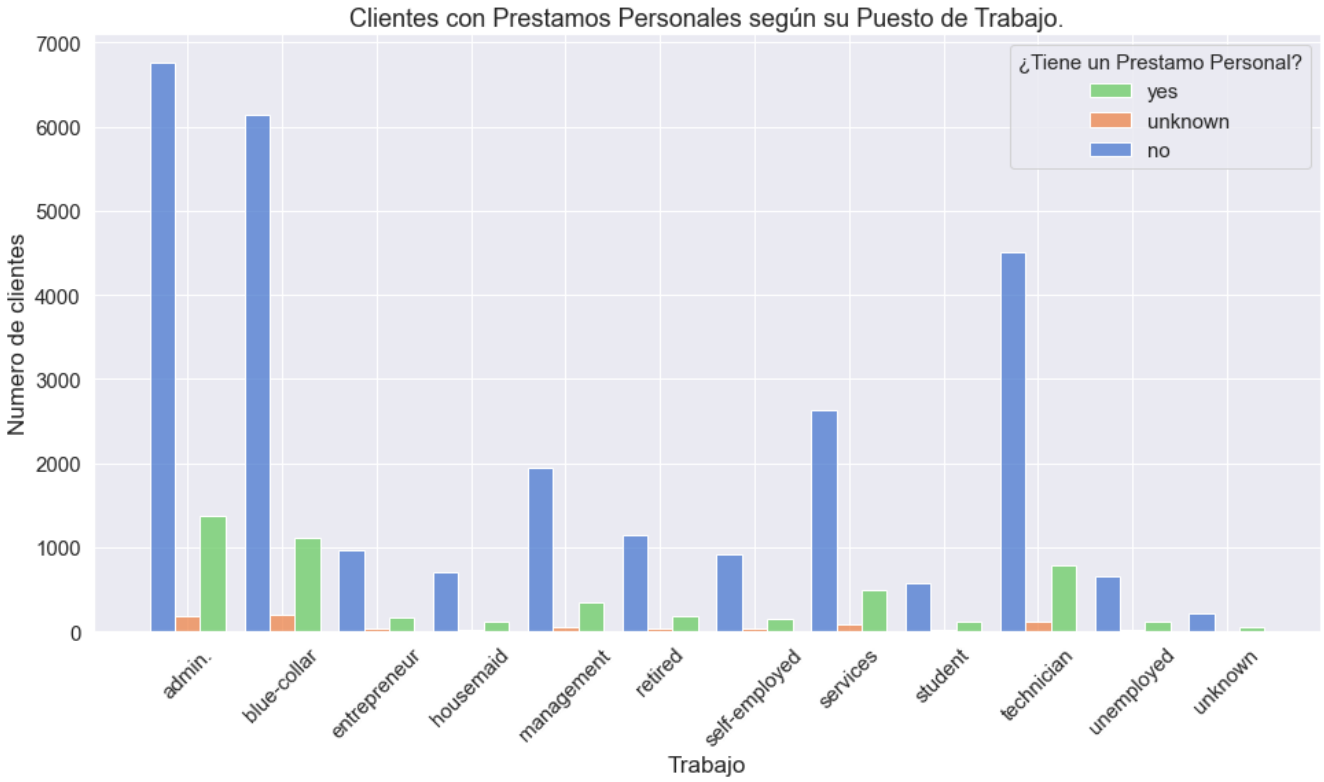

Figura 8-A: Número de clientes con Préstamos Personales según Tipo de Trabajo.

Al analizar la distribución de la población que posee préstamos personales (figura 8-A) podemos ver que la relación que existe entre los grupos (si, no y desconocido) para cada uno de los trabajos es proporcional; es decir que, aunque haya más cantidad de individuos teniendo trabajos administrativos que estudiantes, por ejemplo, la proporción de personas que tienen prestamos personales con respecto a las que no tienen se va a mantener similar.

Al realizar un el mismo tratamiento entre la tenencia de préstamos hipotecarios y el tipo de trabajo de los clientes (figura 8-B) observamos la misma tendencia. Esto nos indica que el puesto de trabajo de cada individuo, en principio, no tendría relación con la tenencia de préstamos de tipo personal o hipotecario.

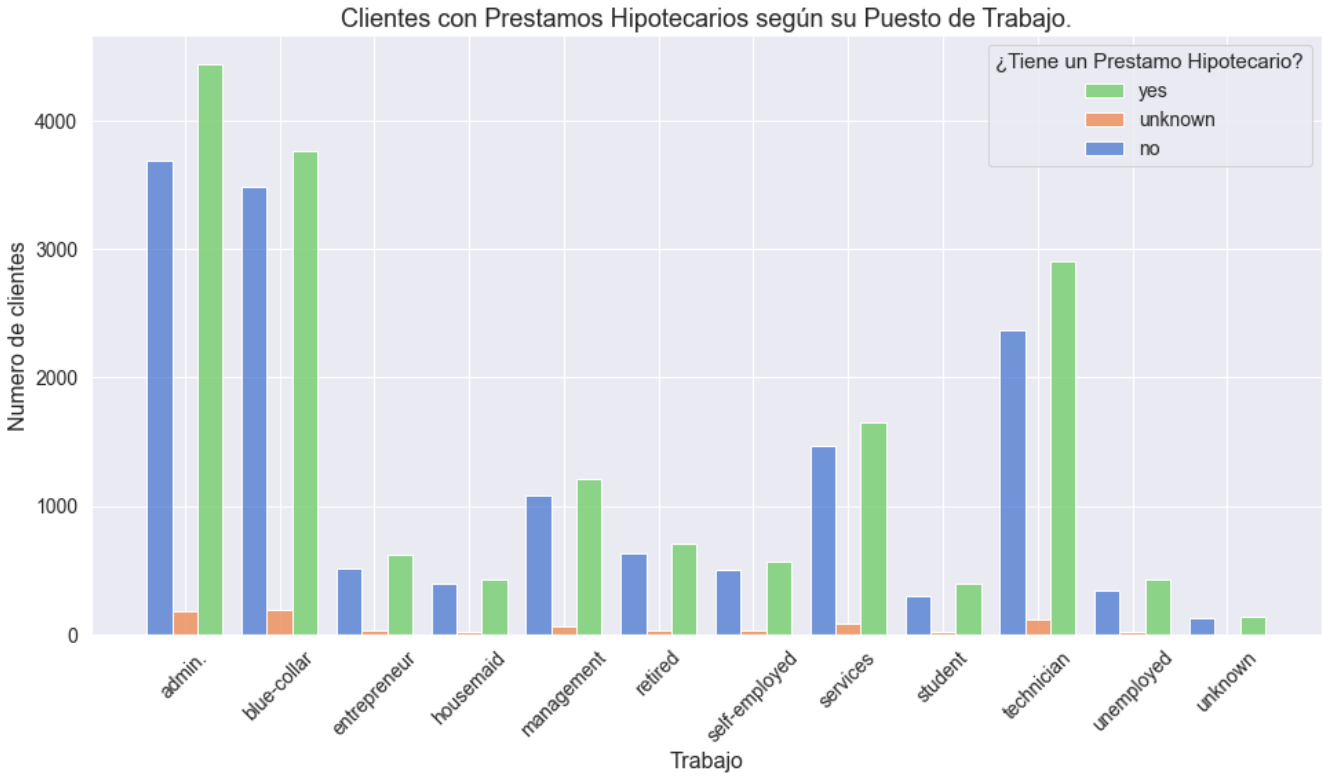

Figura 8-B: Número de Clientes con Préstamos Hipotecarios según Tipo de Trabajo.

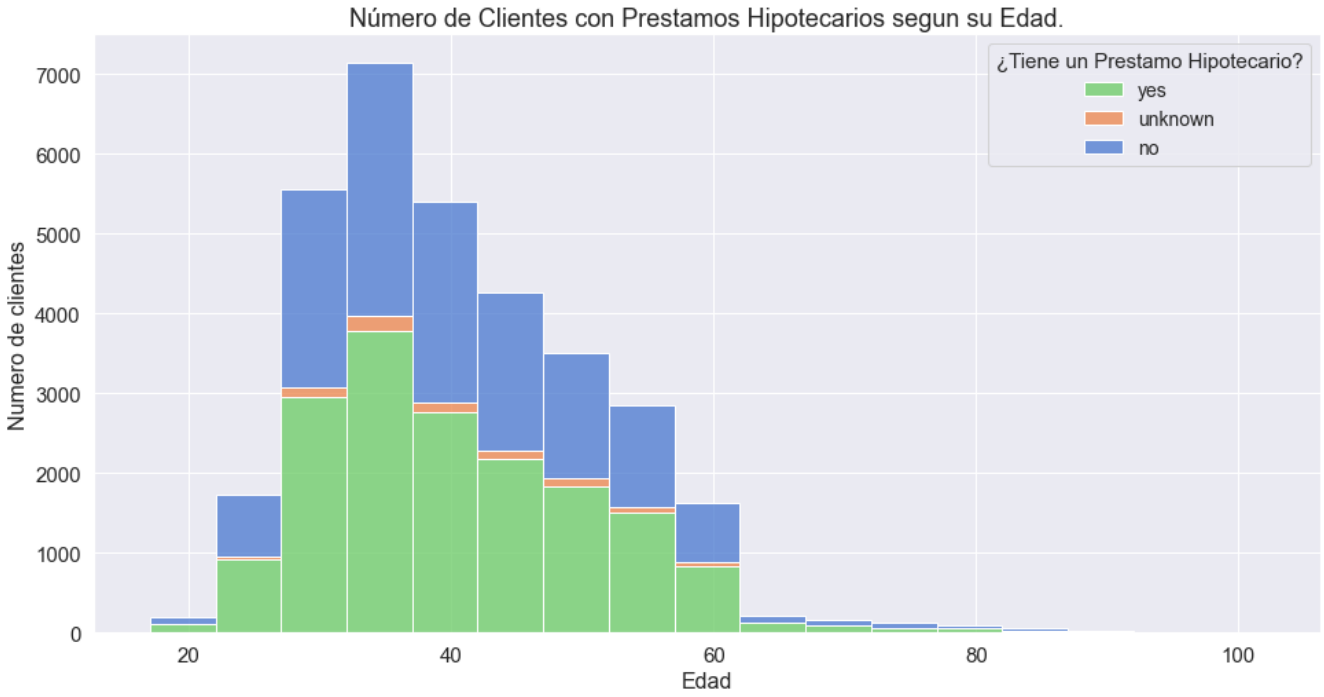

Figura 9-A: Número de Clientes con Préstamos Hipotecarios según su Edad.

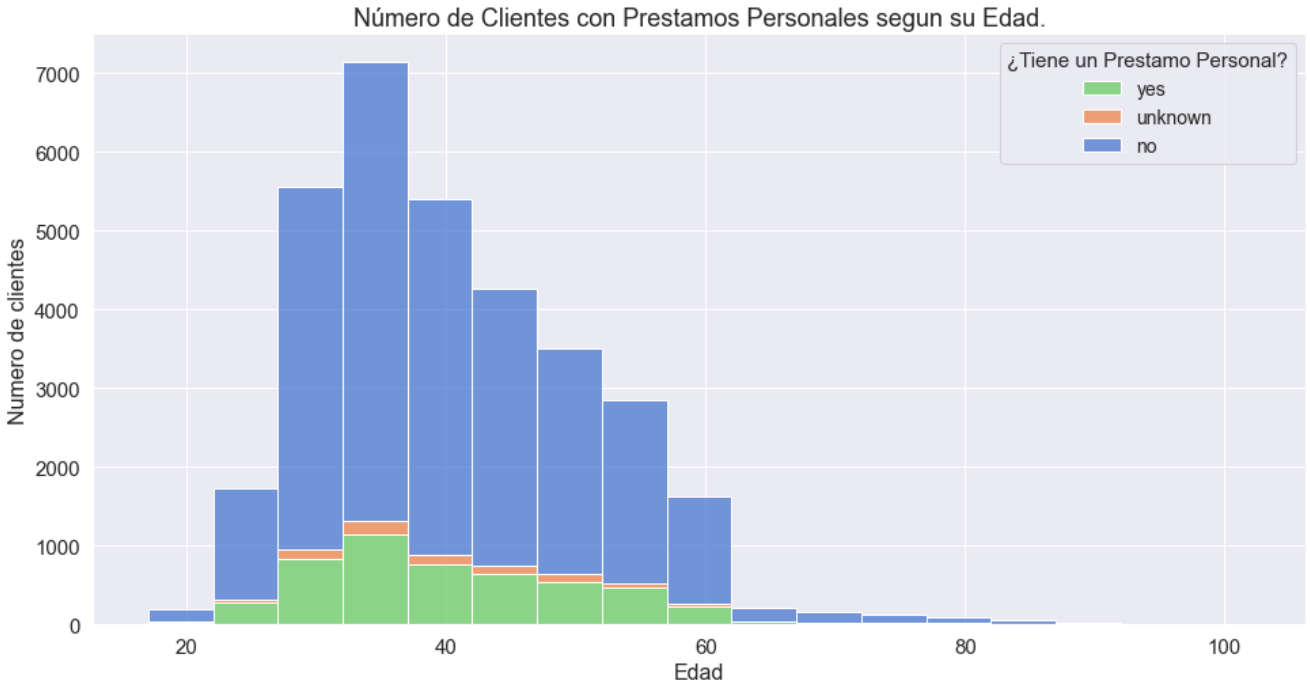

Figura 9-B: Número de Clientes con Préstamos Personales según su Edad.

De la misma forma, si analizamos la relación entre la edad de los individuos y la tenencia de préstamos hipotecarios (figura 9-A) vemos que se repite el mismo patrón que antes: la proporción de gente en cada una de las categorías (si, no y desconocido) se va a mantener y solo dependerá de la cantidad inicial de personas en ese determinado rango de edad. Lo mismo ocurre al analizar la relación entre la edad de los individuos y la tenencia de préstamos personales (figura 9-B).

Teniendo en cuenta todos estos resultados podemos determinar que **la tenencia de prestamos tanto personales como hipotecarios no depende de la edad ni el tipo de trabajo que tengan los clientes del banco**.

Análisis Multivariado.

- Edad de los clientes, Duración del último contacto y Constitución de Plazo Fijo.

El siguiente gráfico (figura 10) se realizó asociando la edad de los clientes y la duración del último contacto telefónico con un empleado del banco.

Se puede observar una diferencia muy marcada entre la población de personas mayores y menores a 60 años: la duración de las llamadas tiende a ser mayor en los clientes de menos de 60 años y mucho menor en personas mayores. Si tenemos en cuenta la distribución de colores en todo el espectro de edades se observa que los clientes que suscribieron a un depósito a largo plazo fueron, en general, quienes mantuvieron llamadas de mayor duración. Adicionalmente, observemos la proporción de personas que constituyeron plazos fijos con respecto a su edad (tabla 2): a pesar de que las personas mayores a 60 años representan un grupo sustancialmente menor de la población, el porcentaje positivo ('yes') es casi cuatro veces mayor que en el caso de las personas menores a 60 años (39,6% vs 10,4%).

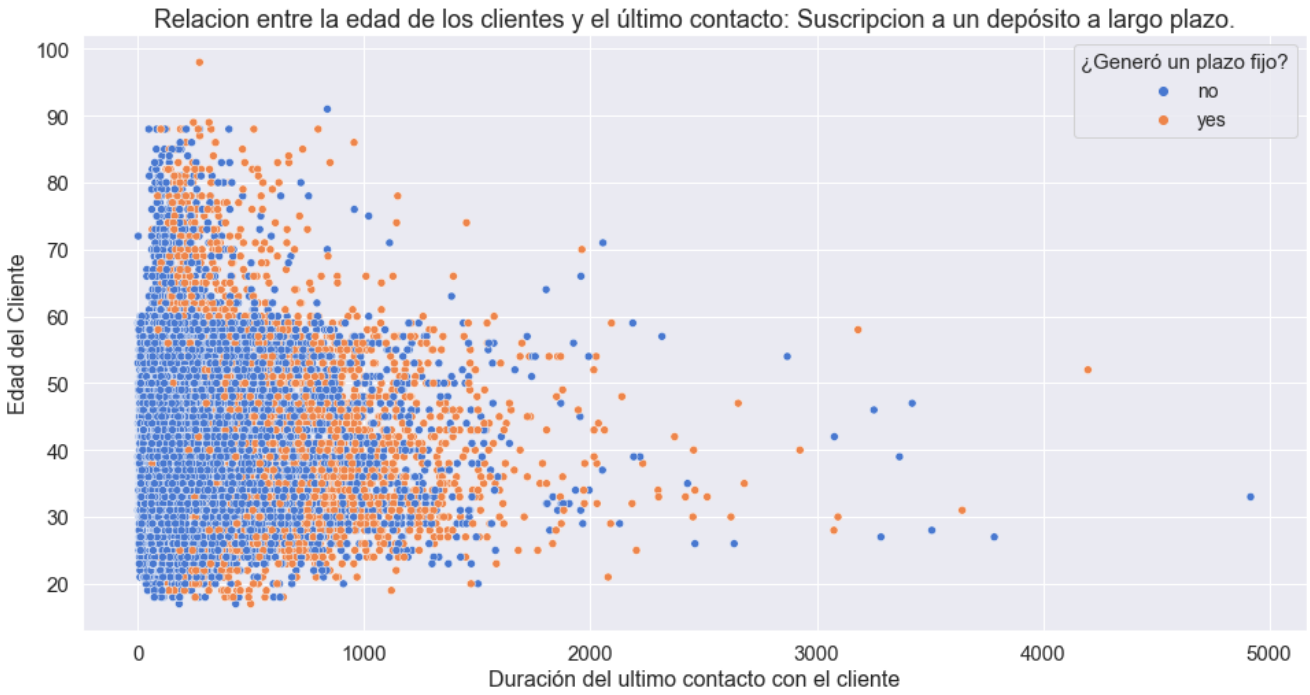

Figura 10: Edad de los clientes, duración del último contacto y constitución de Plazo Fijo.

¿Constituyó un plazo fijo?	Menor de 60 años	Mayor de 60 años
Si	3336 (10,4%)	376 (39,6%)
No	28665 (89,6%)	573 (60,4%)

Tabla 2: Número de personas que constituyeron un plazo fijo con respecto a su edad.

- Tenencia de Préstamos y Constitución de Plazo Fijo.

Continuando con el análisis realizado en la sección 'Análisis Bivariado' (página 12) analizamos la relación entre la tenencia de préstamos personales o hipotecarios y la generación de plazos fijos. Como se puede ver en la Figura 11, al relacionar la cantidad de personas que constituyeron plazos fijos con aquellas que poseen algún tipo de préstamo nuevamente no se observa una tendencia distintiva. Cuando analizamos los números relacionados al grafico sobre préstamos personales podemos ver que la cantidad de personas en cada grupo (si, no o desconocido) es diferente pero que la proporción de personas con o sin plazos fijos se mantiene constante, siendo alrededor del 88,7% negativo y 11,3% positivo. Esto mismo ocurre con respecto a los clientes con préstamos hipotecarios.

¿Constituyó un plazo fijo?	Tiene un préstamo Personal	No tiene un Préstamo Personal
Si	564 (11,2%)	3058 (11,3%)
No	4459 (88,8%)	24073 (88,7%)

Tabla 3: Relación entre Clientes con Préstamos Personales y Constitución de Plazos Fijos.

¿Constituyó un plazo fijo?	Tiene un préstamo Hipotecario	No tiene un Préstamo Hipotecario
Si	1994 (11,6%)	1628 (10,9%)
No	15260 (88,4%)	13272 (89,1%)

Tabla 4: Relación entre Clientes con Préstamos Hipotecarios y Constitución de Plazos Fijos.

Figure 11: Tenencia de Préstamos Hipotecarios o Personales y Generación de un Plazo Fijo.

Considerando todos estos resultados podemos determinar que no existe una relación entre la generación de plazos fijos y la tenencia de prestamos hipotecarios o personales, la edad o el tipo de trabajo del individuo por lo que podemos dejar estos datos en segundo plano y concentrarnos en otros frentes para encontrar estrategias de marketing aprovechables.

- Circunstancias del último contacto y constitución de plazos fijos.

Analizamos la relación entre la longitud media del último contacto entre empleado y cliente según el mes en que se realizó, y nos propusimos ver si presenta alguna relación con la suscripción a un plazo fijo o depósito a largo plazo.

Si vemos la figura 12-A en conjunto con la figura 10, nuevamente se observa que para los casos positivos (barras naranjas) la duración de las llamadas resulto significativamente más larga que en los casos negativos (barras azules). Sin embargo, en los meses de marzo, septiembre y octubre la duración en los casos "positivos" fue sustancialmente más corta. El mes de diciembre será tomado como un caso intermedio: a pesar de que la longitud de las llamadas es un poco más corta que en los casos de mayo a agosto y noviembre, la desviación estándar es mucho más grande que en otros meses.

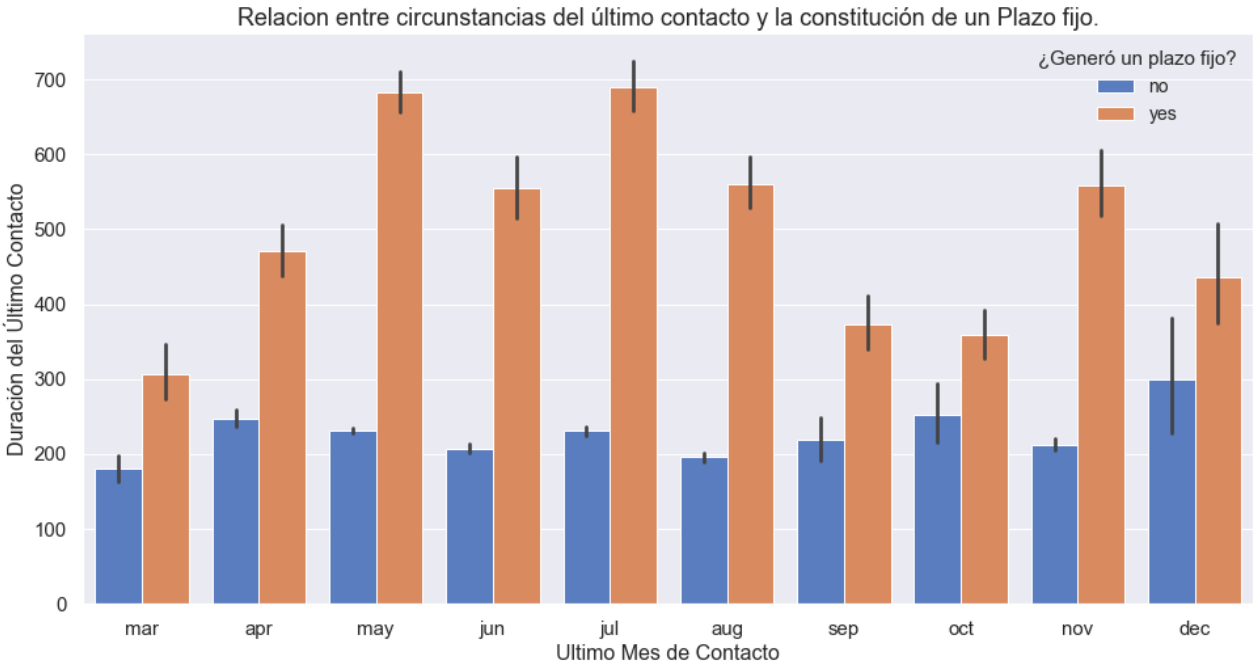

Figura 12-A: Relación entre las circunstancias del último contacto y la constitución de un Plazo Fijo.

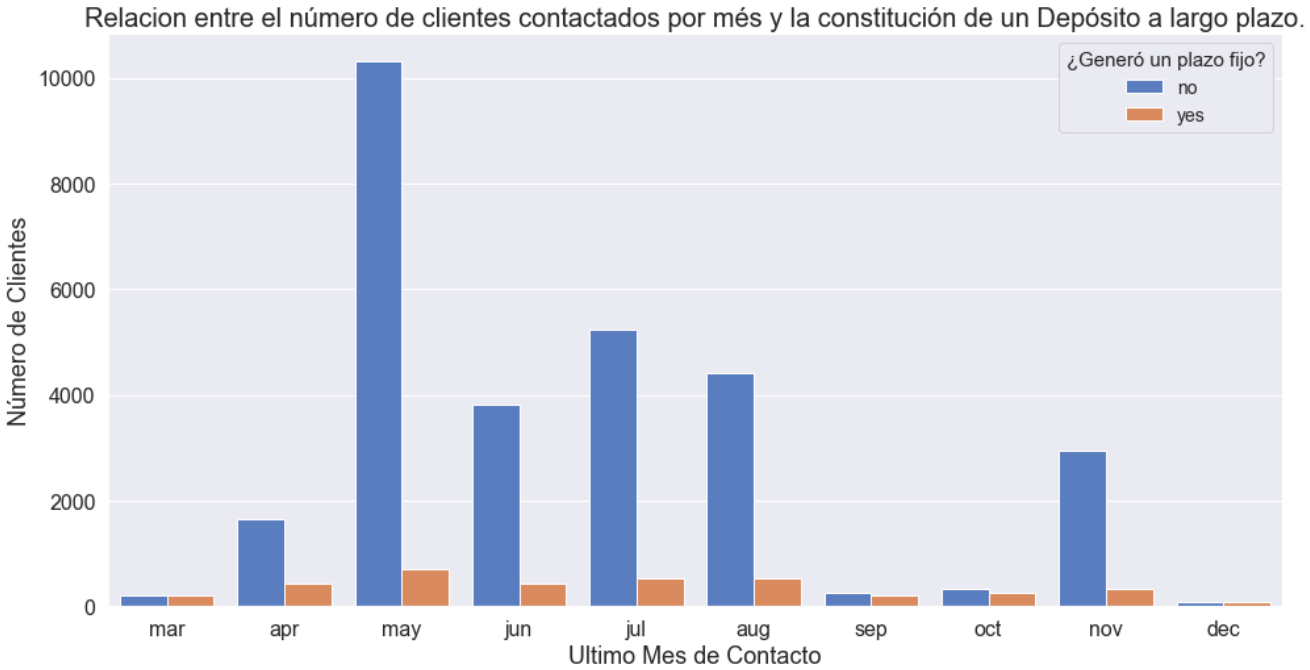

Figura 12-B: Relación entre el Número de Clientes contactados por última vez para cada mes y la constitución de un Plazo Fijo.

Por otro lado, graficamos la cantidad de individuos que fueron contactados por última vez en cada mes y si constituyeron o no plazos fijos (figura 12-B), y observamos que de mayo a agosto y también en noviembre, momentos donde la duración de las llamadas fue mayor para los casos positivos, también se tuvieron la menor proporción de casos positivos con respecto al total. Por el otro lado, en marzo, septiembre, octubre y diciembre, donde se observa que la longitud de las llamadas es mucho menor para los casos positivos (nuevamente considerar el caso de diciembre con precaución), aunque la cantidad de individuos contactados es sustancialmente menor a la de los otros meses el porcentaje de casos positivos con respecto al total se acerca a casi la mitad.

- Edad de los clientes, último contacto y constitución de un Plazo Fijo.

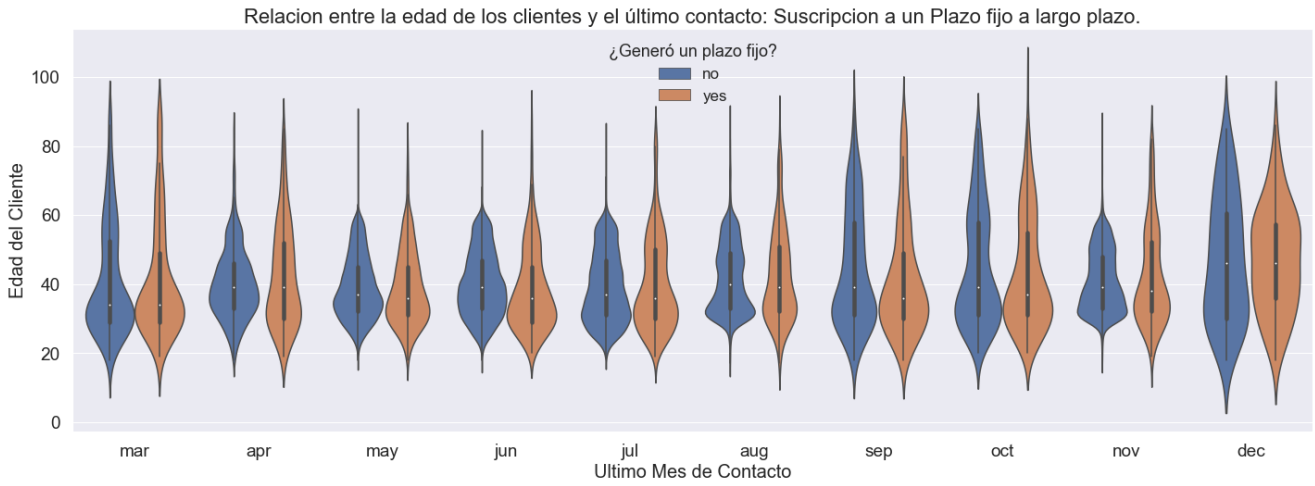

Figura 13: Relación entre la edad de los clientes del banco, el mes del último contacto y la constitución de un Plazo Fijo.

Al graficarse la relación entre la edad de los clientes y el último mes de contacto, y considerando los resultados previos, se puede observar que en los meses de interés (marzo, septiembre, octubre y diciembre) la cantidad de clientes mayores a 60 años aumenta sustancialmente, viéndose como un ensanchamiento en el segmento superior del gráfico de violín.

Hipótesis: posible estrategia de márketing.

Teniendo en cuenta todos los resultados previos podemos definir la siguiente hipótesis: **en los meses de interés (marzo, septiembre, octubre y diciembre) la cantidad de contactos con individuos mayores a 60 años es mayor que en el resto de los meses y la duración de las llamadas en el caso de estos individuos es mucho menor que con los clientes menores a 60 años**. Una posible estrategia de marketing si es que esta hipótesis se mantiene (lo que requiere el análisis de la actividad de otros conjuntos de clientes o la incorporación de más registros a nuestra base de datos) podría **ser la presentación de ofertas o beneficios para mayores de 60 años en estos meses**.

Selección del Algoritmo.

En la siguiente sección se mostrará el análisis realizado sobre diferentes algoritmos de clasificación, cuales fueron nuestros parámetros de selección del mejor modelo y los pasos realizados para la optimización del este.

Entrenamiento y selección del modelo de clasificación.

- Selección de los modelos.

Se utilizaron los modelos provistos en dos bibliotecas, PyCaret y Scikit-Learn. Se realizó la comparación de los modelos provistos por la biblioteca PyCaret usando la base de datos con todas las variables y la variable 'y' como blanco. Al analizar las matrices de confusión de cada uno de los modelos nos centraremos en el número de verdaderos y falsos positivos detectados, dado que no solo queremos predecir la mayor cantidad de resultados positivos posibles (clientes que constituyen un plazo fijo) sino que queremos evitar pérdidas económicas por la utilización de recursos en clientes que finalmente no constituirían depósitos a largo plazo.

- PyCaret.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.9039	0.8794	0.3497	0.6569	0.4556	0.4083	0.4332	0.3790
lr	Logistic Regression	0.9023	0.8748	0.3310	0.6511	0.4380	0.3908	0.4181	0.8320
catboost	CatBoost Classifier	0.9023	0.8762	0.3490	0.6400	0.4511	0.4026	0.4252	4.3140
lightgbm	Light Gradient Boosting Machine	0.9022	0.8744	0.3610	0.6332	0.4591	0.4098	0.4298	0.0830
ada	Ada Boost Classifier	0.9021	0.8757	0.3358	0.6454	0.4415	0.3937	0.4192	0.1640
xgboost	Extreme Gradient Boosting	0.9017	0.8735	0.3441	0.6377	0.4464	0.3977	0.4209	0.4920
lda	Linear Discriminant Analysis	0.9005	0.8745	0.3922	0.6063	0.4758	0.4237	0.4363	0.0270
ridge	Ridge Classifier	0.8991	0.0000	0.2824	0.6413	0.3917	0.3458	0.3808	0.0160
svm	SVM - Linear Kernel	0.8962	0.0000	0.2313	0.6331	0.3311	0.2898	0.3353	0.0350
knn	K Neighbors Classifier	0.8958	0.7888	0.3742	0.5741	0.4527	0.3980	0.4093	0.2800
et	Extra Trees Classifier	0.8907	0.7384	0.3302	0.5436	0.4102	0.3540	0.3677	0.3360
dt	Decision Tree Classifier	0.8906	0.6894	0.3249	0.5440	0.4064	0.3503	0.3647	0.0200
rf	Random Forest Classifier	0.8891	0.7952	0.3550	0.5289	0.4244	0.3657	0.3748	0.4490
dummy	Dummy Classifier	0.8847	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0100
nb	Naive Bayes	0.8316	0.8312	0.4223	0.3348	0.3647	0.2710	0.2778	0.0150

qda	Quadratic Discriminant Analysis	0.1922	0.0000	0.9000	0.1038	0.1861	0.0000	0.0000	0.0190

Tabla 5: Comparación de puntajes para los diferentes modelos de la biblioteca PyCaret.

Se seleccionó el modelo 'Light Gradient Boosting Machine' por tener el conjunto de puntajes de Accuracy (correspondiente al número de predicciones correctas), Recall (proporción de verdaderos positivos correctos) y Precision (proporción de verdaderos positivos con respecto al total de positivos) más balanceado. Se realizó el entrenamiento del modelo definiendo el peso de las variables como balanceado (class_weight="balanced"), con el cual nos aseguramos de que para cada subconjunto de entrenamiento cada clase tenga un peso inversamente proporcional a la cantidad de registros que tenemos: para la clase 1 (resultado positivo, el cliente creó un plazo fijo), como tenemos menor cantidad de registros le dará mayor peso a la hora de entrenar el modelo.

Luego de entrenar nuestro modelo se utilizó para predecir los valores de la variable blanco (a los que llamaremos valores predichos) y luego se compararon los resultados con los valores de la variable 'y' verdaderos. Para hacer esta comparación se graficó la matriz de confusión. Posteriormente se realizó el ajuste del modelo usando el comando 'tune_model' y se procedió de la misma forma. Los resultados obtenidos fueron:

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Mean	0.8596	0.9270	0.8582	0.4351	0.5773	0.5038	0.5457
Std	0.0071	0.0079	0.0203	0.0143	0.0148	0.0179	0.0173

Tabla 6: Puntajes del entrenamiento del modelo 'Light Gradient Boosting Classifier'

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Mean	0.7326	0.8592	0.8855	0.2798	0.4250	0.3078	0.3947
Std	0.0185	0.0152	0.0284	0.0140	0.0171	0.0220	0.0215

Tabla 7: Puntajes del entrenamiento del modelo 'Light Gradient Boosting Classifier' ajustado.

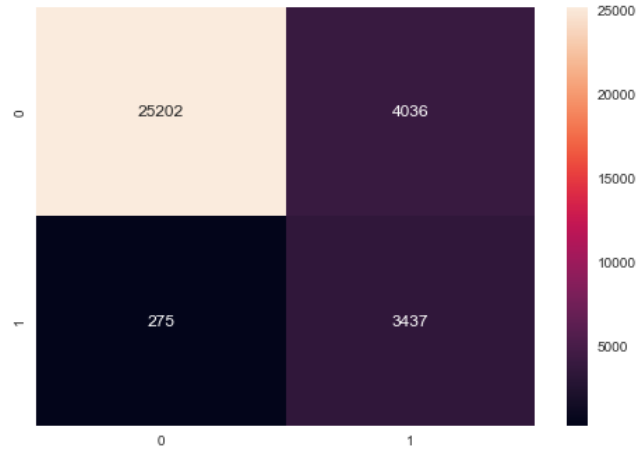

Figura 15: Matriz de confusión para el modelo 'Light Gradient Boosting Classifier'.

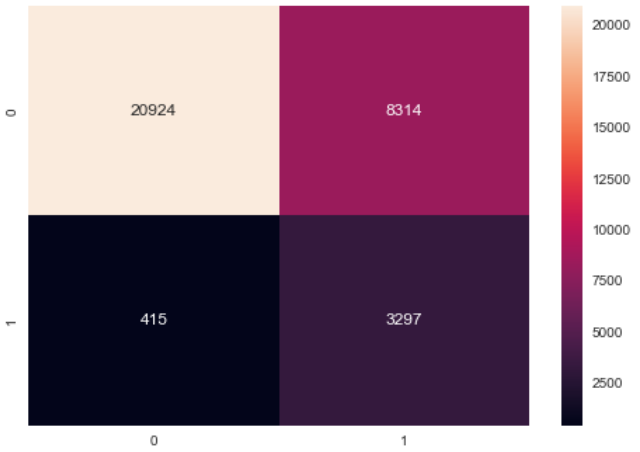

Figura 16: Matriz de confusión para el modelo 'Light Gradient Boosting Classifier' ajustado.

Como vemos el comparar la media de los puntajes para el modelo inicial y ajustado y las matrices de confusión, los resultados obtenidos son sustancialmente mejores con respecto a los verdaderos positivos y los falsos positivos para el modelo sin ajustar. Seleccionaremos entonces el modelo sin ajustar para continuar con el análisis.

- Scikit-Learn.

Se realizó el entrenamiento de los modelos 'Decision Tree Classifier', 'Random Forest Classifier', 'Logistic Regresion' y 'K-Neighbors Classifier'. Inicialmente se separaron los datos en 6 subconjuntos de entrenamiento y testeo, se ajustaron manualmente algunos de los parámetros de dichos modelos y se calcularon las puntuaciones de Accuracy y Recall promedio para los subconjuntos.

Al igual que con el modelo definido con la biblioteca PyCaret, analizamos las matrices de confusión de cada uno de los modelos de Scikit-Learn.

Modelo	Parámetros Ajustados	Subconjunto	Recall	Accuracy
Decision Tree Classifier	Profundidad Máxima (max_depth): 5, Conjunto aleatorio usado (random_state): 0, Peso de cada clase (class_weight): "balanced"	Entrenamiento	0,870	0,783
		Testeo	0,860	0,780
Random Forest Classifier	Profundidad Máxima (max_depth): 6, Peso de cada clase (class_weight): "balanced"	Entrenamiento	0,836	0,837
		Testeo	0,816	0,832
Logistic Regression	Peso de cada clase (class_weight): "balanced"	Entrenamiento	0,753	Nan
		Testeo	0,755	Nan
K-Neighbors Classifier	Normalizado (StandatrScaler, LabelEncoder), Número de vecinos (n_neighbors): 5	Entrenamiento	0.476	0.921
		Testeo	0.380	0.899

Tabla 8: Modelos entrenados, parámetros definidos manualmente y puntajes obtenidos.

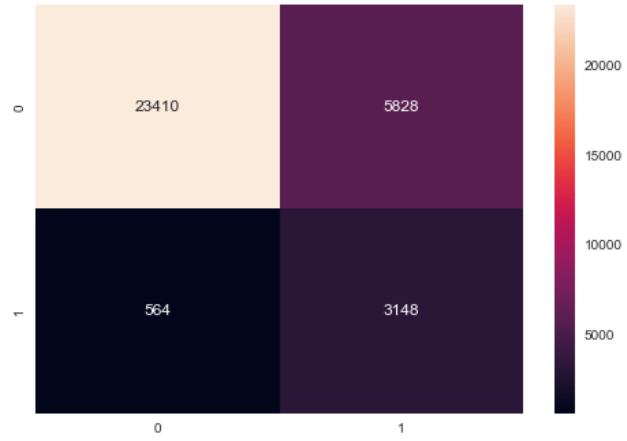

Figura 17: Matriz de confusión (Decision Tree Cassifier).

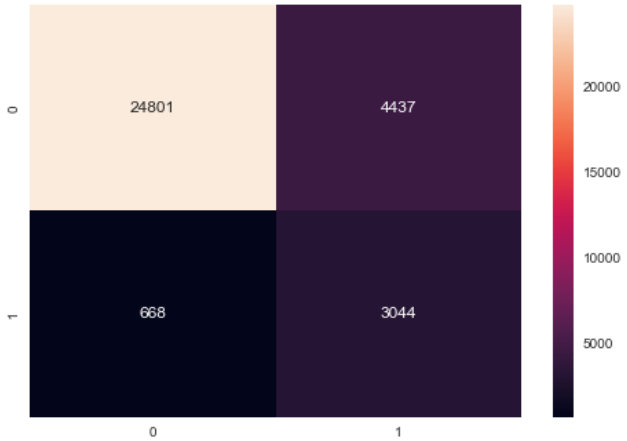

Figura 18: Matriz de confusión (Random Forest Cassifier)

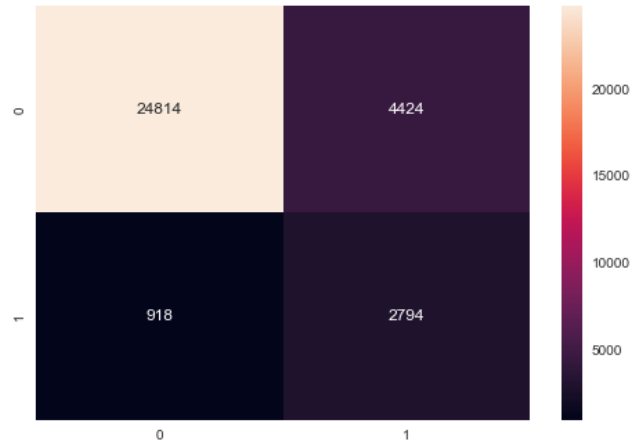

Figura 19: Matriz de confusión (Logistic Regresion)

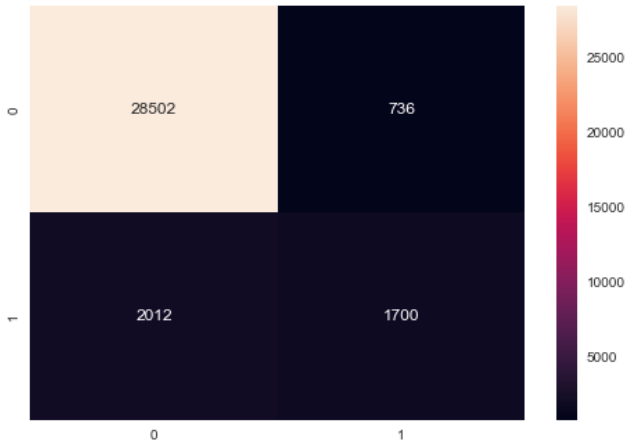

Figura 20: Matriz de confusión (K-Neighbors Classifier)

Al comparar los resultados obtenidos descubrimos un resultado inesperado: a pesar de que los puntajes obtenidos para el modelo 'Light Gradient Boosting Machine' fueron las mejores, seguido por el resto de los modelos y por último las del modelo 'K-Neighbors Classifier', las matrices de confusión nos indican que este último tiene el menor porcentaje de falsos positivos de todos (un orden de magnitud más chico que el resto). Aunque la cantidad de verdaderos positivos es más chica (aproximadamente la mitad que la del resto de los modelos) y la cantidad de falsos negativos es mucho mayor (un orden de magnitud mayor), nos resulta más importante centrarnos en el número de falsos y verdaderos positivos para seleccionar nuestro mejor modelo: **no solo queremos obtener las mayores ganancias posibles (dependiente de los verdaderos positivos) sino que queremos evitar las pérdidas económicas relacionadas al uso equivocado de nuestros recursos**, enfocándonos por error en personas que no tengan intenciones de establecer plazos fijos.

Teniendo en cuenta estos resultados decidimos quedarnos con el modelo 'K-Neighbors Classifier'. Sin embargo, sería importante mejorar nuestro modelo para disminuir el número de falsos negativos.

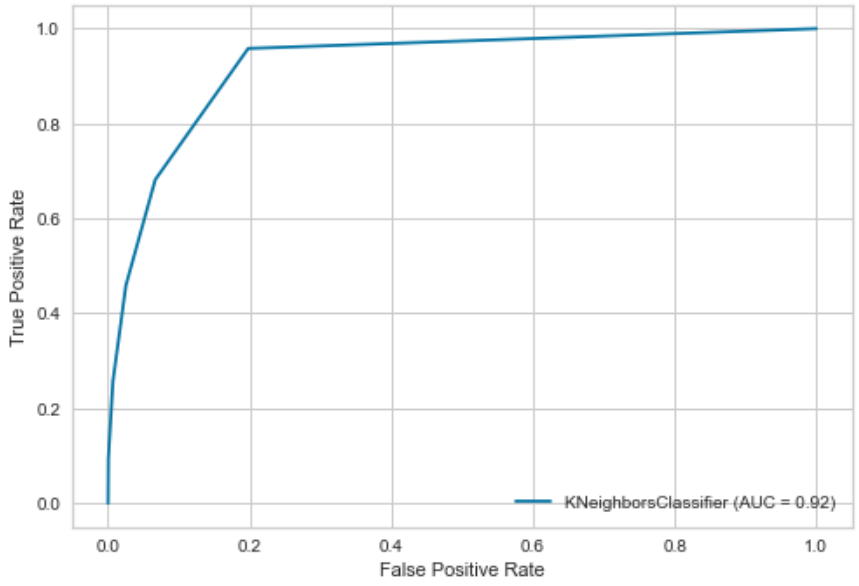

Figura 21: Curva ROC del modelo 'K-Neighbors Classifier'.

Optimización del modelo

Como se menciono anteriormente, para realizar el mejoramiento de los modelos utilizados previamente se utilizó la biblioteca BorutaShap. Debido a las limitaciones que esta tiene (solo permite su utilización en modelos basados en árboles de decisión) no se pudo utilizar para seleccionar variables en los modelos 'K-Neighbors Classifier' ni 'Logistic Regression', y tampoco podemos extrapolar los resultados obtenidos en el resto de los modelos. Por otro lado, se utilizaron los métodos 'RandomizedSearchCV' y 'GridSearchCV' para seleccionar los mejores valores de los parámetros del modelo 'K-Neighbours Classifier'.

- Selección de variables.

Como se dijo anteriormente, la biblioteca BorutaShap solo puede ser utilizada en modelos basados en árboles de decisión. Para no descartar alguno de los modelos entrenados anteriormente se decidió optimizar las variables seleccionadas y reentrenar los modelos para tratar de obtener mejores resultados. En cada sección se compararán las matrices de confusión obtenidas.

- Light Gradient Boosting Machine:

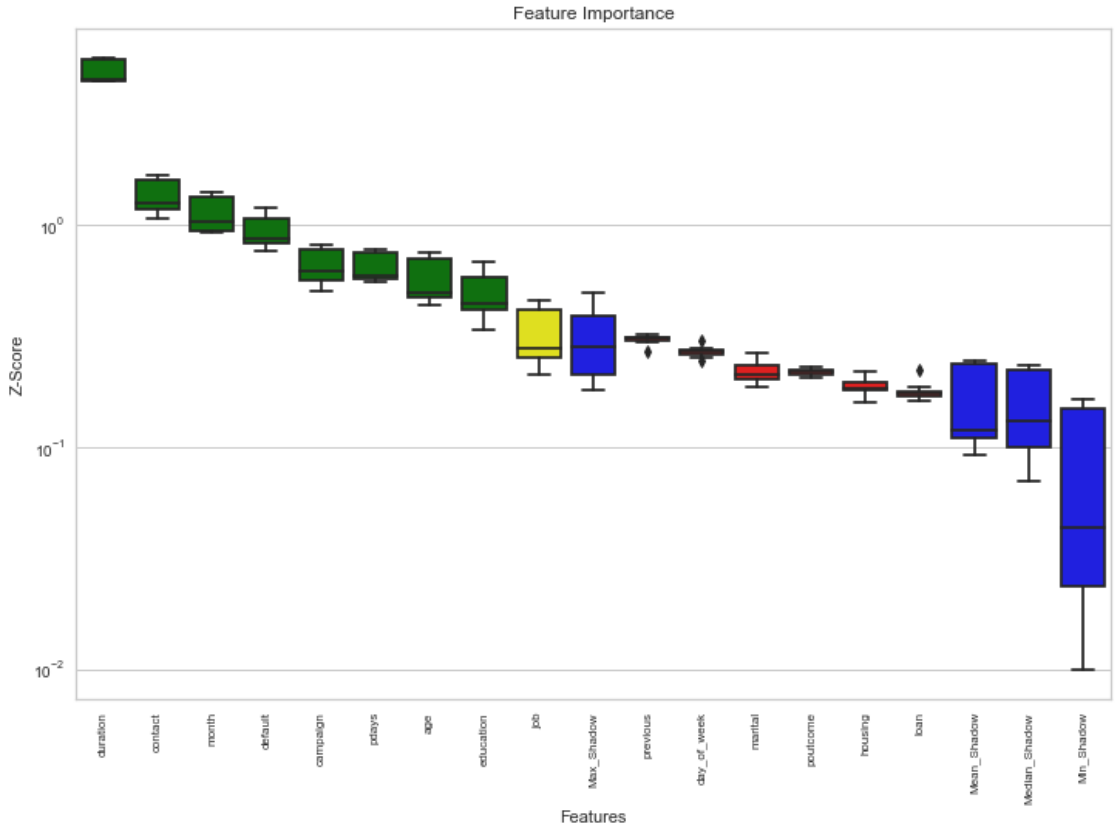

Figura 22: Resultados de Boruta Shap para el modelo 'Light Gradient Boosting Machine'

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Mean	0.8529	0.9269	0.8631	0.4240	0.5685	0.4921	0.5374
Std	0.0077	0.0054	0.0156	0.0148	0.0144	0.0176	0.0158

Tabla 9: Puntajes del entrenamiento del modelo 'Light Gradient Boosting Machine' (Boruta Shap).

Las variables seleccionadas como aceptadas son 'campaign', 'month', 'duration', 'age', 'pdays', 'contact' y 'default', mientras que la seleccionada como tentativa fue 'poutcome'.

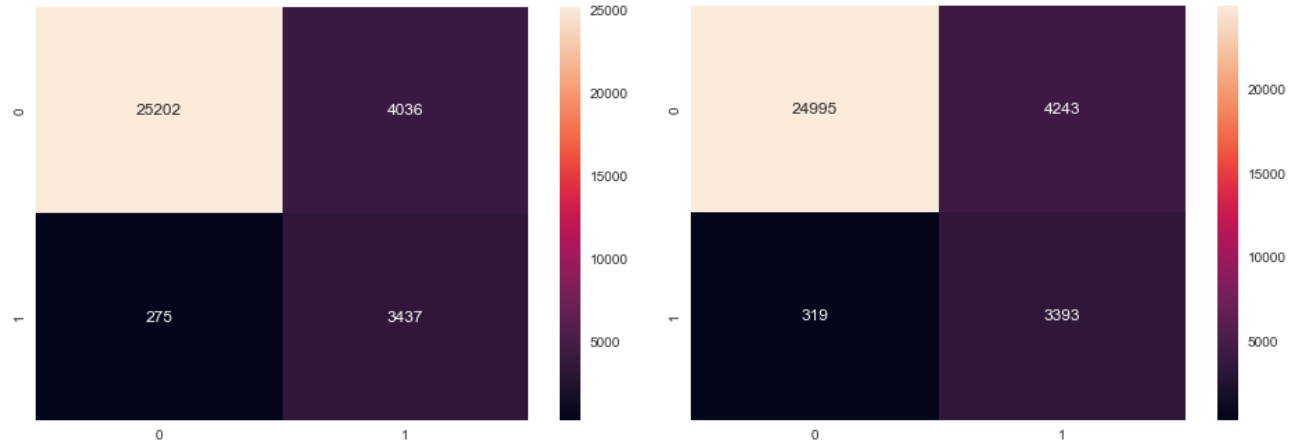

Figura 23: Comparación de matrices de confusión para el modelo original (izquierda) y el nuevo (derecha).

Como vemos, no se produjo una mejora en los resultados obtenidos.

- Decision Tree Classifier.

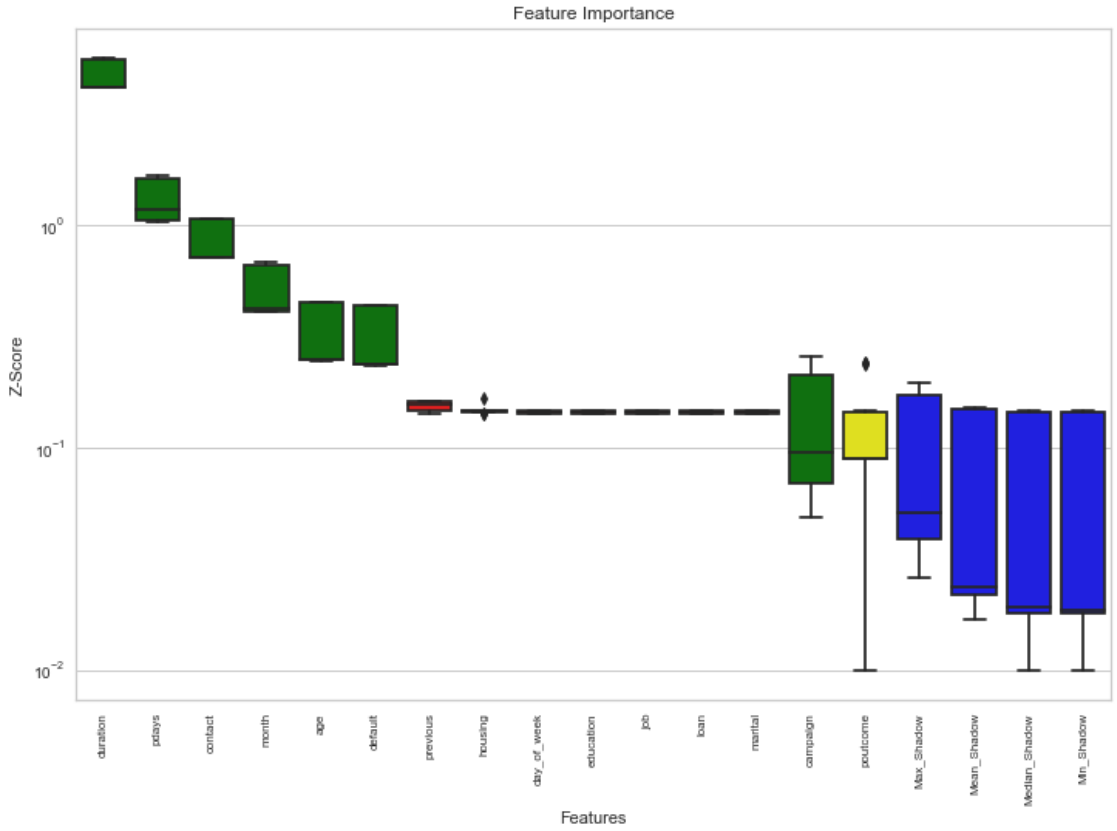

Figura 24: Resultados de Boruta Shap para el modelo 'Decision Tree Classifier'.

Subconjunto	Recall	Accuracy
Entrenamiento	0.870	0.783
Testeo	0.860	0.780

Tabla 10: Puntajes del entrenamiento del modelo 'Decision Tree Classifier' (Boruta Shap).

Las variables seleccionadas como aceptadas son 'campaign', 'month', 'duration', 'age', 'pdays', 'contact' y 'default' mientras que la seleccionada como tentativa fue ‘poutcome’.

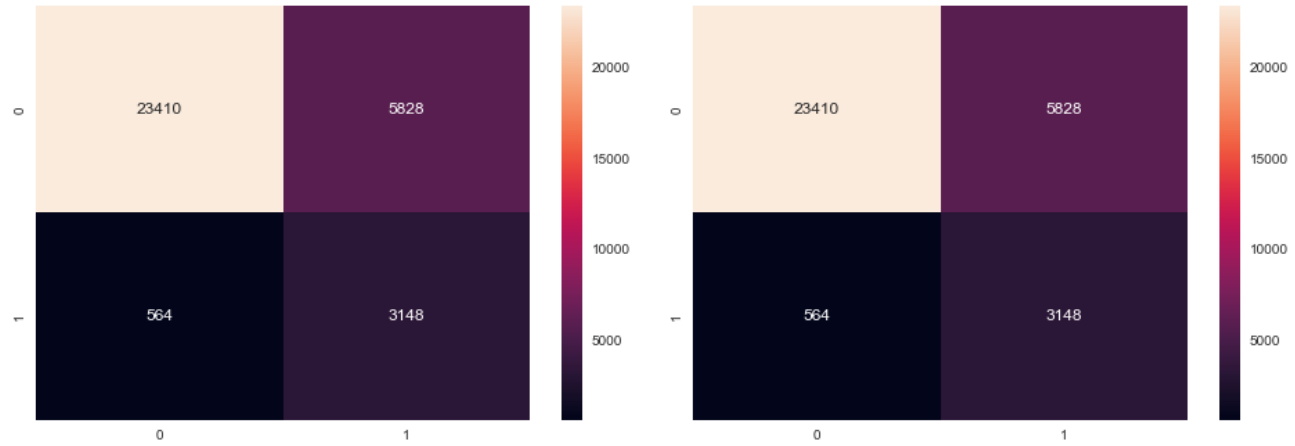

Figura 25: Comparación de matrices de confusión para el modelo original (izquierda) y el nuevo (derecha).

Como vemos, los resultados obtenidos son iguales.

- Random Forest Classifier.

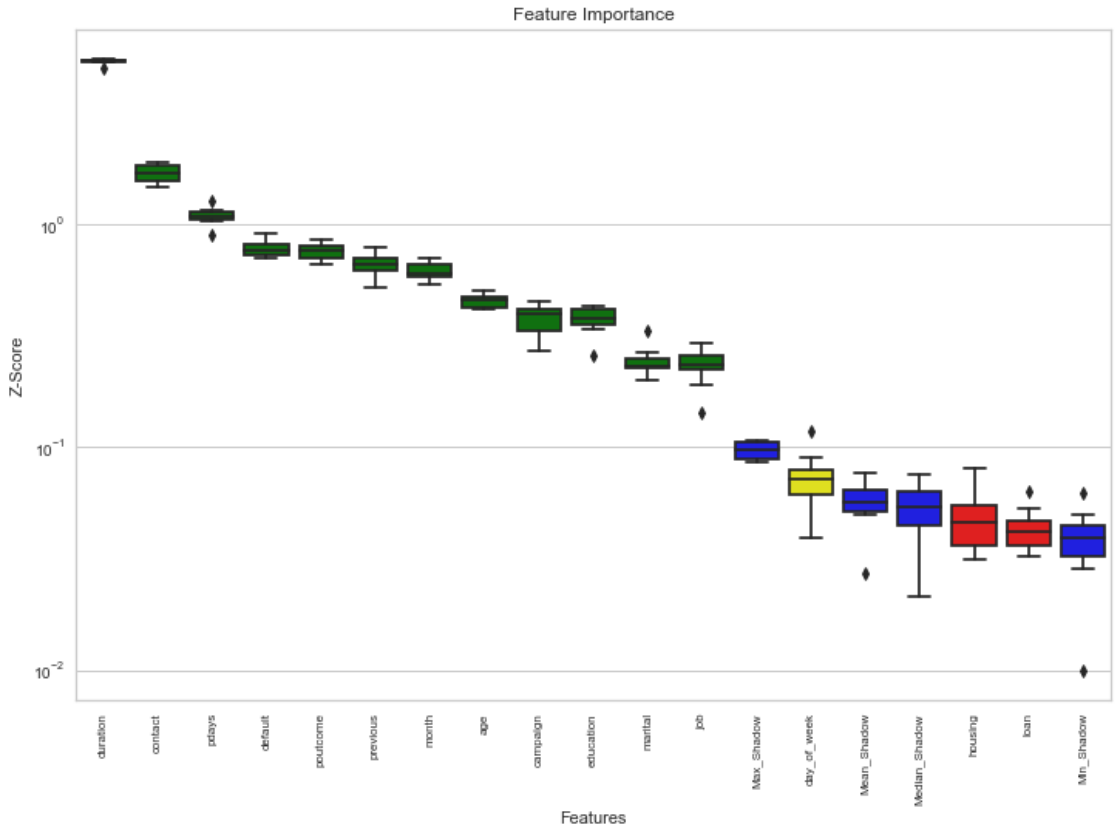

Figura 26: Resultados de Boruta Shap para el modelo 'Random Forest Classifier'.

- Ajuste de Hiperparámetros.

Método	Parámetros Ajustados
RandomizedSearchCV	Número de Vecinos ('n_neighbors'): 28, Número de Puntos en cada Nodo ('leaf_size'): 24
GridSearchCV	Número de Vecinos ('n_neighbors'): 9, Número de Puntos en cada Nodo ('leaf_size'): 2

Tabla 11: Valores obtenidos con el ajuste de hiperparámetros.

COMENTARIO: los valores de los puntos para la selección aleatoria (RandomizedSearchCV) y en grilla (GridSearchCV) son diferentes debido a que los mecanismos de la selección en grilla requieren demasiada capacidad de cómputo, de la cual no disponemos suficiente para profundizar más el análisis y tuvimos que reducir sustancialmente el número de puntos a comprobar.

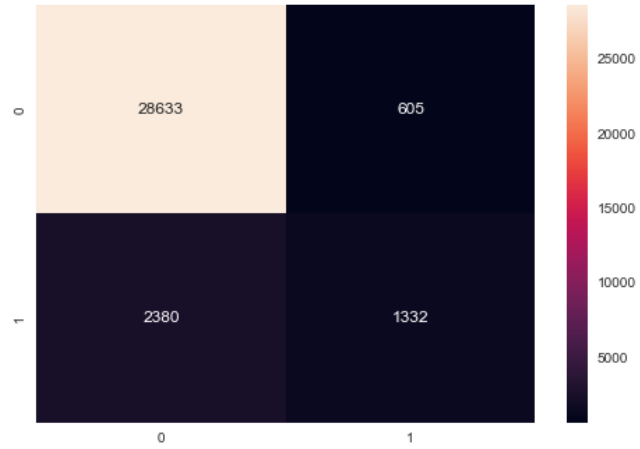

Figura 26: Matriz de confusión (Modelo mejorado 'KNN' con RandomSearchCV).

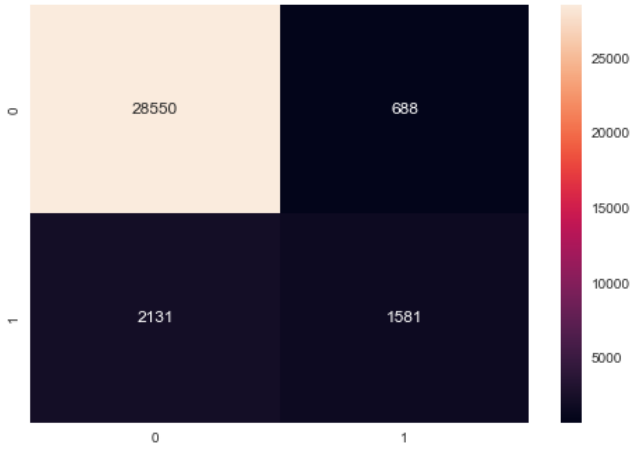

Figura 27: Matriz de confusión (Modelo mejorado 'KNN' con GridSearchCV).

Si comparamos los resultados obtenidos en las matrices de confusión con el resultado del modelo sin ajustar se puede ver que este último tiene mejores números de verdaderos y falsos positivos, por lo que decidimos quedarnos con el primero realizado.

Resultado Final y Conclusiones

Considerando los resultados obtenidos luego del trabajo de mejoramiento realizado decidimos mantener el modelo 'K-Neighbors Classifier'. A pesar de que tenemos una menor proporción de verdaderos positivos con respecto a los otros modelos, la cantidad de falsos positivos es la menor entre todos los grupos. La elección se tomó luego de sopesar los resultados obtenidos: aunque no obtengamos la mayor cantidad de clientes posible que acepten definir un plazo fijo (tiene la menor proporción de verdaderos positivos de todo el grupo) no vamos a tener una pérdida de recursos tan grande como con el resto de los modelos al enfocarnos en los clientes equivocados (tiene la menor cantidad de falsos positivos). Sin embargo, se deberá tratar de mejorar el modelo para disminuir el número de falsos negativos.

Si se sigue manteniendo el comportamiento observado en la subpoblación correspondiente a mayores de 60 años (constitución de plazos fijos con duración de llamadas más cortas, en especial en los meses de marzo, septiembre, octubre y diciembre) también podríamos ofrecer promociones especiales que logren que más gente decida constituir plazos fijos.

Para mejorar nuestro modelo podríamos intentar los siguientes pasos:

- tratar de obtener una mayor cantidad de registros o tratar de obtener una mejor capacidad de cómputo que nos permita revisar en mayor profundidad los datos de que disponemos, dado que una de nuestras mayores limitantes es el hardware.
- La codificación de las variables categóricas no es la óptima, dado que se usaron números consecutivos (proceso de codificación autimática).

Nuevamente mostraremos los resultados obtenidos para el modelo seleccionado:

Modelo: K-Neighbors Classifier.

Parámetros	Número de Puntos en cada Nodo ('leaf_size'): 30, Número de Vecinos ('n_neighbors'): 5, Parámetro de la Métrica de Minowski ('p'): 2, Datos normalizados.
Puntajes	Recall (entrenamiento): 0.48, Recall (testeo): 0.38, Accuracy (entrenamiento): 0.92, Accuracy (testeo): 0.90

Tabla 12: Tabla resumen.

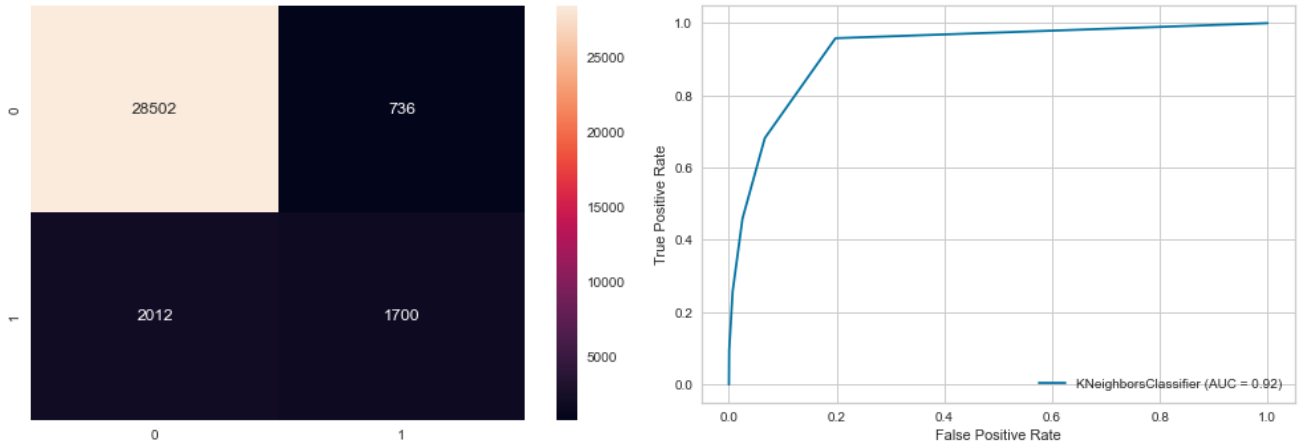

Figura 28: Matriz de confusión y Curva ROC del modelo seleccionado.