92.26 - Seminario de Ingeniería Industrial IV

MACHINE LEARNING

Cuatrimestre: 2°C 2021

Integrantes:

Nombre y apellido	Padrón
Luciano Castro	96853
Ignacio Ravazzini	97045
Paula Gonzalez Sandoval de Herrera	98363
Juan Ignacio Berta	102505
Juan Cruz Camacho	102376

Informe final

Introducción

Para comenzar el trabajo importamos el dataset a un Google Colab Notebook. Una vez que lo leímos utilizando la librería Pandas, comenzamos a investigarlo para conocer las características y variables del mismo.

Como primera medida, buscamos la dimensión del dataset y vimos que contenía 17 columnas (es decir, 17 variables) y 45211 filas (es decir, 45211 observaciones).

Luego, como se pedía en la consigna, cambiamos los valores de la variable respuesta Y de 2: Si - 1: No a 1: Si, 0: No. Además, analizamos la distribución de la variable respuesta: 39922 personas no acceden a plazo fijo y 5289 personas si lo hacen. Vemos que según los datos históricos, la mayoría de la gente no accede al plazo fijo (clases desbalanceadas).

Como siguiente paso, pasamos a responder las preguntas del punto 1:

- De las 17 variables tenemos 7 que son numéricas y 10 categorías:
  Las numéricas son Age, Balance (euros), Last contact day, Last contact duration, Campaign, Pdays y Previous. Mientras que Job, Marital Status, Education, Credit, Housing Loan, Personal Loan, Contact, Last Contact Month, Poutcome y Subscription.

- En cuanto a los valores vacíos, no hay NA's pero si hay información faltante en las variables categóricas simbolizada con un "unknown".
  En las variables Job y Education la frecuencia de "unknown" no parece tener tanta incidencia. En la variable Contact tenemos un tercio de las observaciones con información faltante. Además, en la variable Poutcome (resultado de la campaña anterior) hay un 80% de observaciones faltantes. Esto sucede porque la mayoría de los clientes analizados no participaron de campañas anteriores. Consecuentemente, la variable Pdays también tiene una gran cantidad de valores faltantes simbolizados con un "-1", que se corresponden con las personas que no participaron de las campañas anteriores.

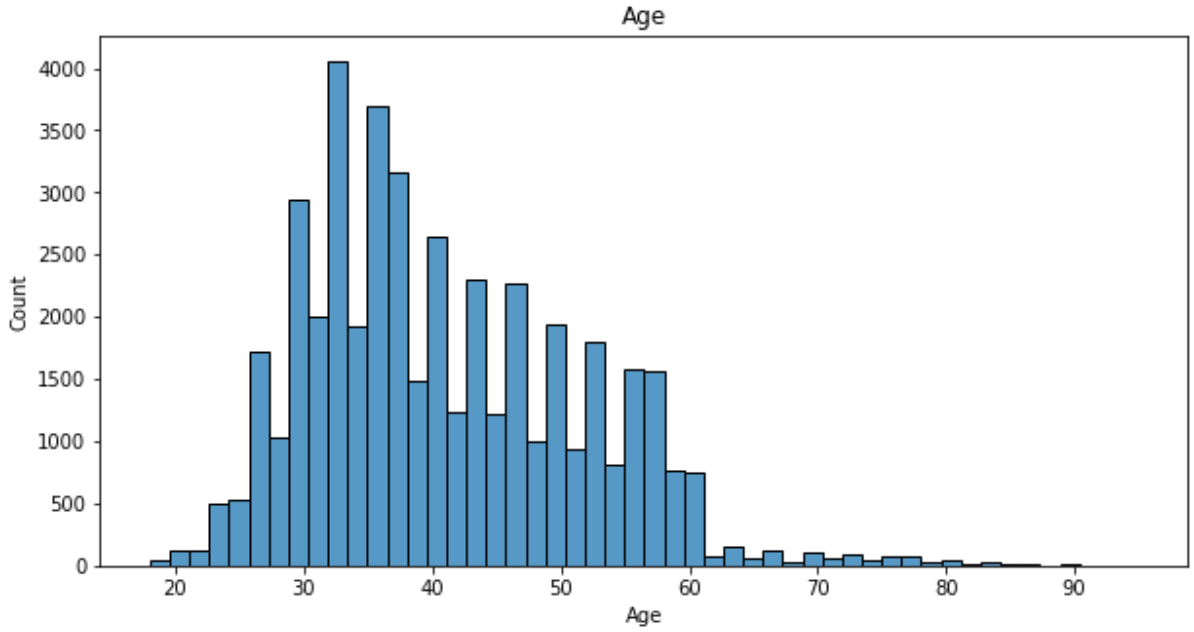
- La variable Marital Status se distribuye de la siguiente manera: 60,19% son casados, 28,29% están solteros y el 11,51% están divorciados

- La variable Education se distribuye de la siguiente manera: 51,32% tienen educación secundaria, 29,42% tienen educación terciaria, 15,15% tienen educación primaria y el 4,11% restante corresponde a la categoría "unknown" (información faltante)

- Se suscribieron al plazo fijo 5289 personas (11,7%), mientras que 39922 (88,3%) no lo hicieron

Análisis y visualización de variables

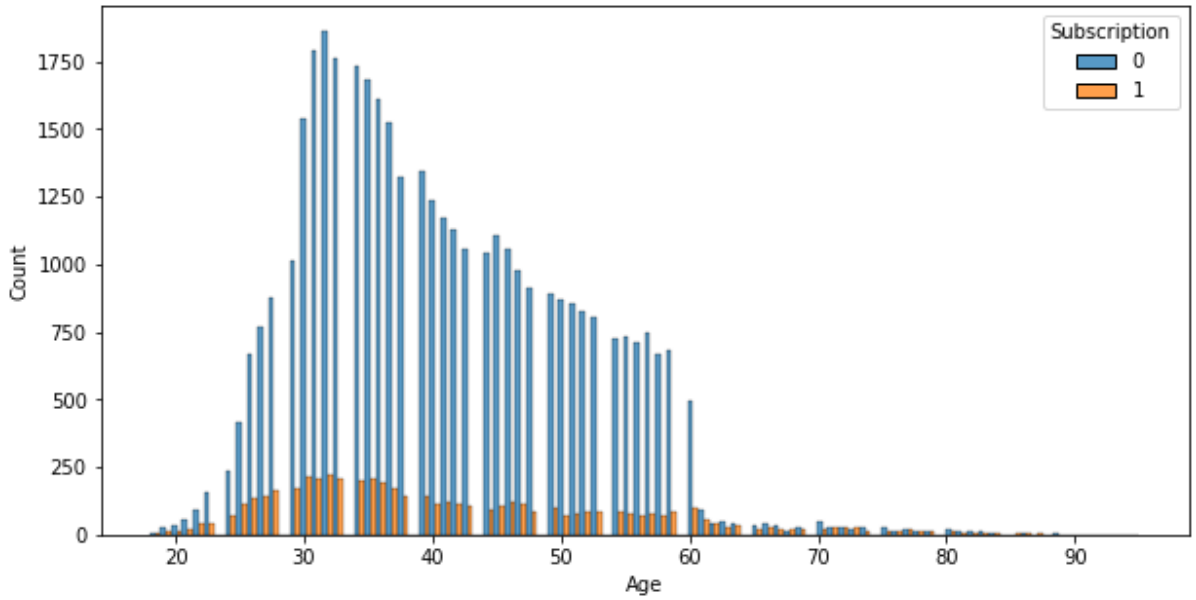
Age: Vemos que existe un cambio en la distribución a partir de los 60 años.

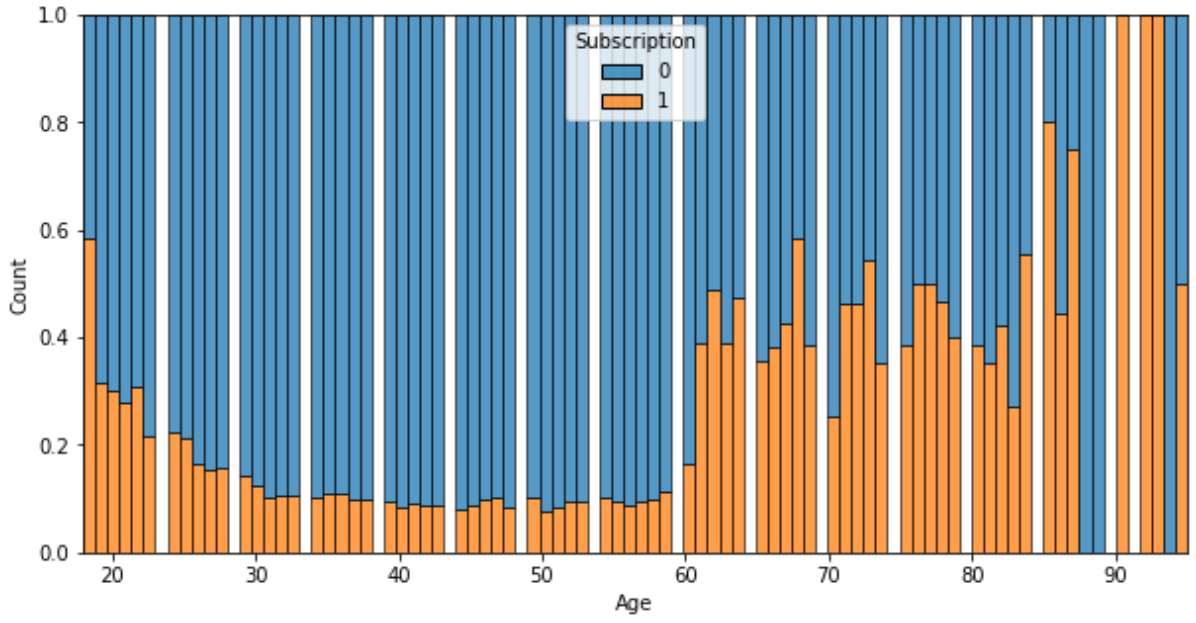


También notamos que a partir de los 60 años cambia la proporción de la variable respuesta

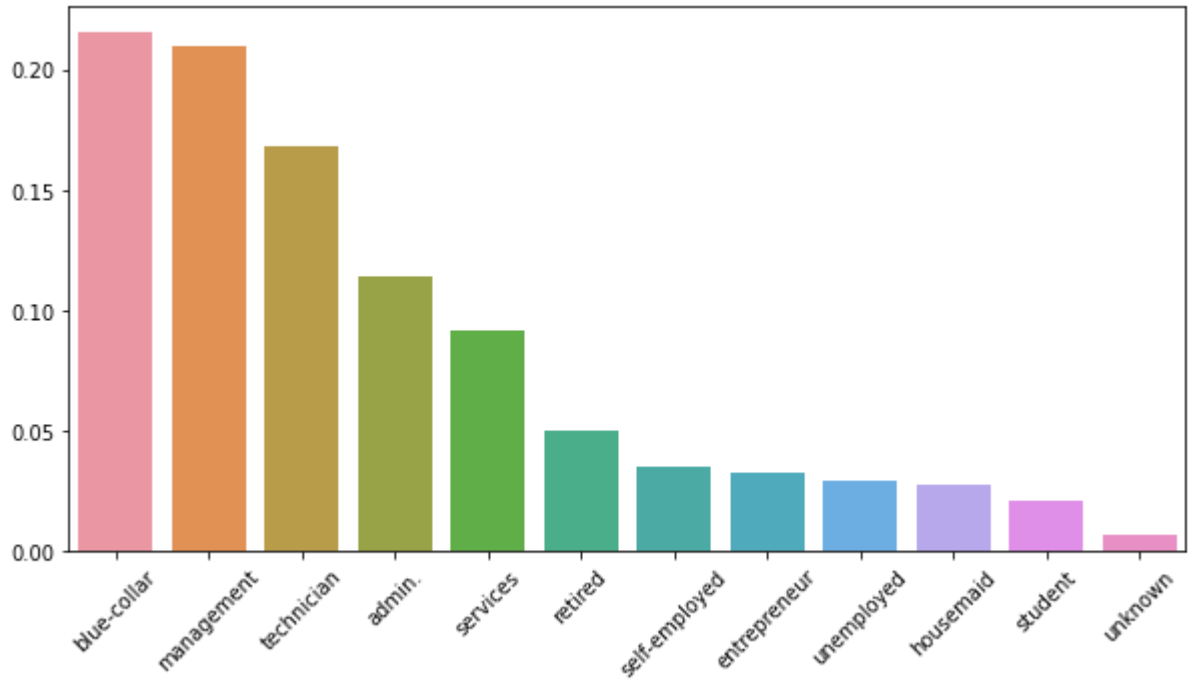
Subscription	Age	Mayor60	Menor60
0		0.891261	0.577441
1		0.108739	0.422559

Los resultados de las proporciones anteriores se da por el corte abrupto que tiene la variable Age.

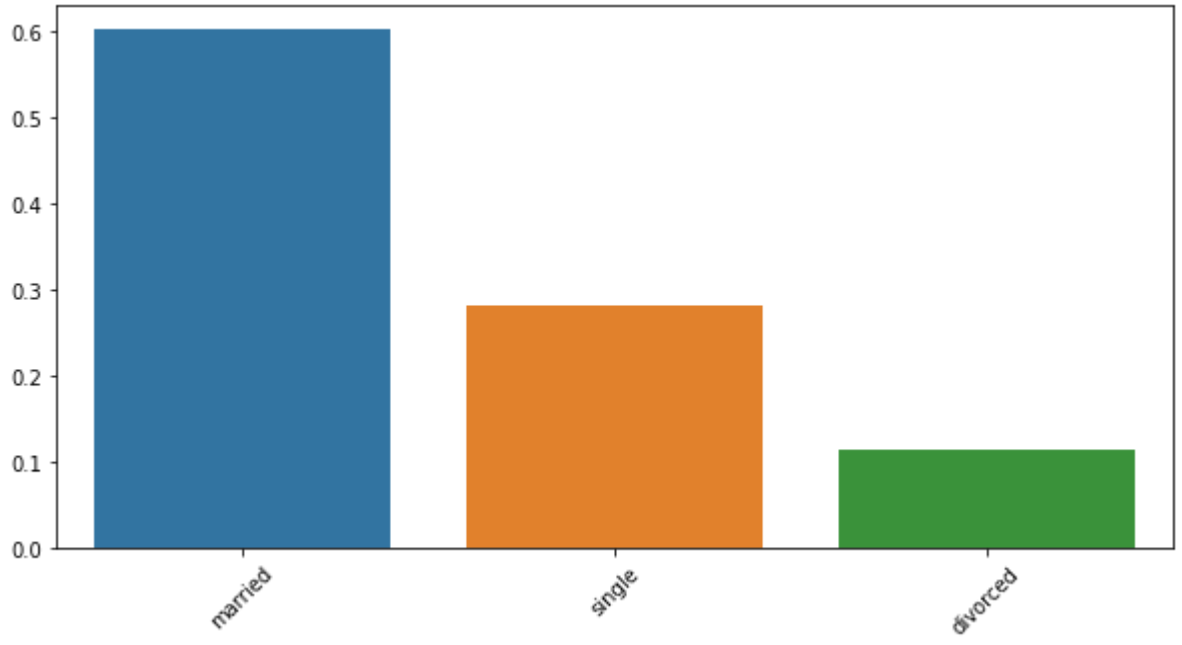




Job: En esta variable, visto que había varias categorías con muy pocas observaciones (menos del 5%), decidimos agruparlas en una nueva categoría llamada “Others”. Estas categorías con baja frecuencia son: 'self-employed', 'entrepreneur', 'unemployed', 'housemaid', 'student', 'unknown'.



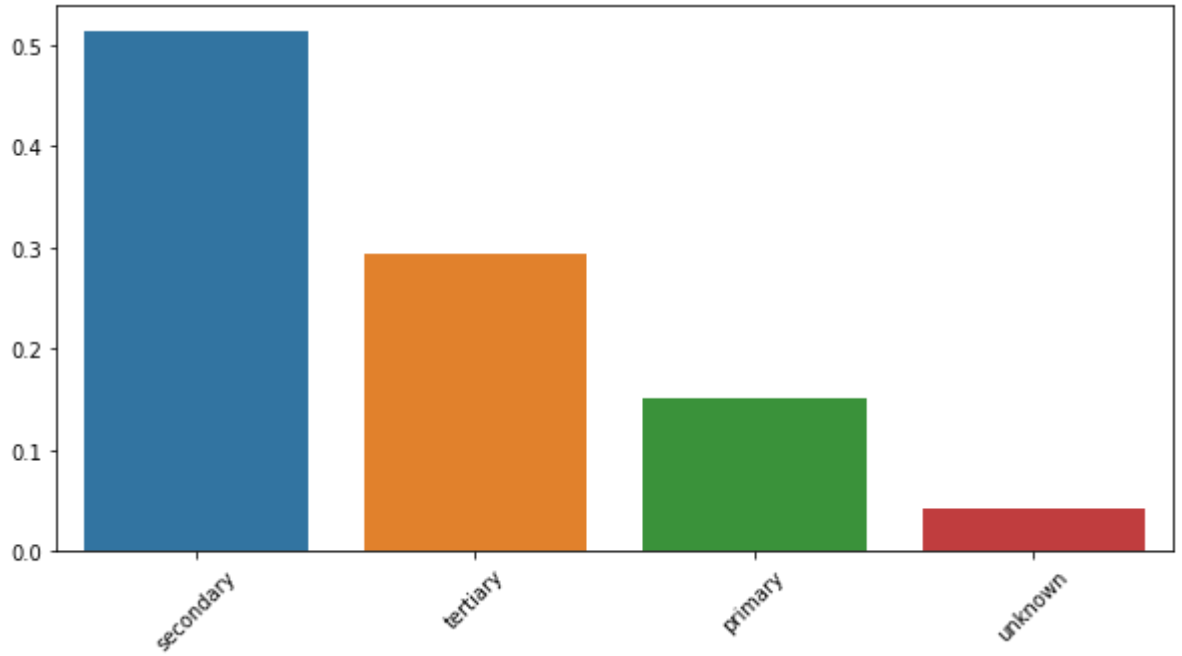
Marital Status: No hay grandes diferencias entre las distintas clases. Además, vemos que la la personas solteras son más propensas a acceder al plazo fijo



Marital Status	divorced	married	single
Subscription			
0	4585	24459	10878
1	622	2755	1912

Marital Status	divorced	married	single
Subscription			
0	0.880545	0.898765	0.850508
1	0.119455	0.101235	0.149492

Education: Existen valores faltantes pero resultan menos del 5% del set. Imputamos los datos faltantes con la categoría de mayor frecuencia por tratarse de pocas observaciones.

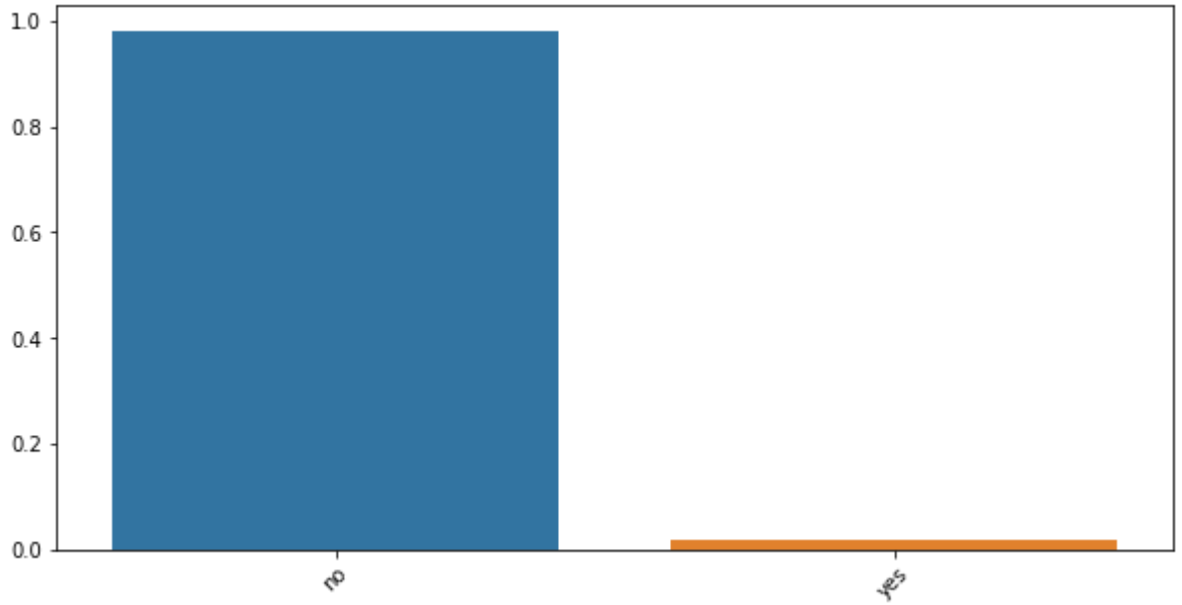


Además vemos que a mayor nivel educativo, mayor acceso al plazo fijo.

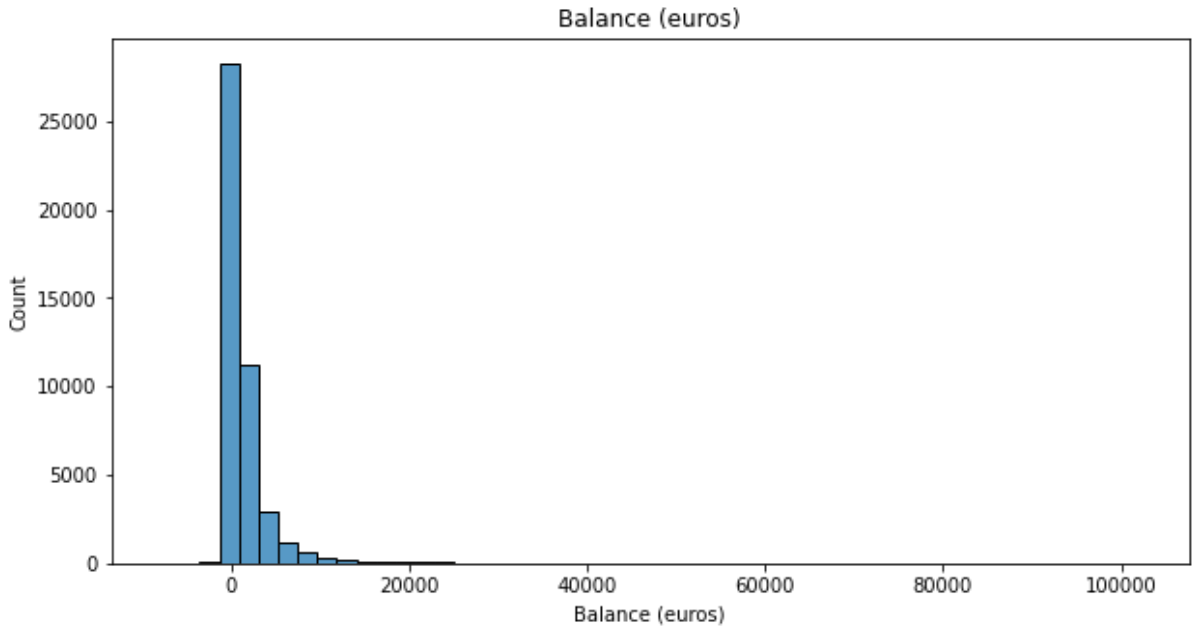
Education	primary	secondary	tertiary	unknown
Subscription				
0	6260	20752	11305	1605
1	591	2450	1996	252

Education	primary	secondary	tertiary	unknown
Subscription				
0	0.913735	0.894406	0.849936	0.864297
1	0.086265	0.105594	0.150064	0.135703

Credit: Tiene una categoría con muy baja frecuencia, por lo que la mejor opción es excluirla del análisis.



Balance (euros): Es una variable con distribución asimétrica y con valores negativos

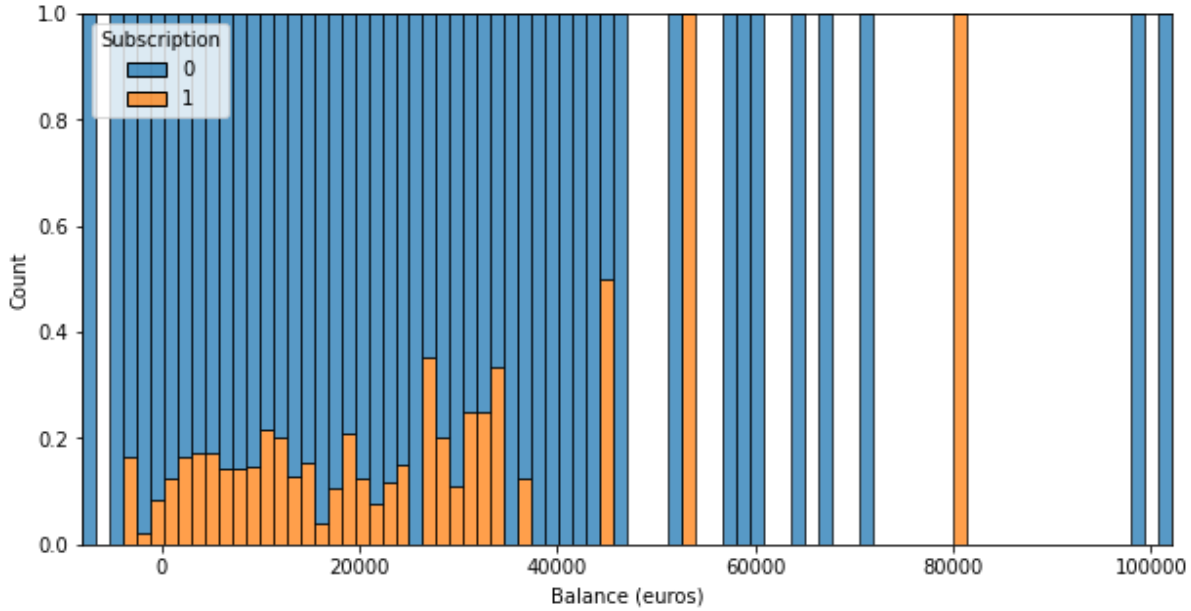


Si partimos la distribución en cuantiles de 10% podemos ver que a medida que subimos el cuantil, tenemos un aumento de la proporción de éxito en la campaña. Las personas que tienen un menor balance acceden menos al plazo fijo.

Balance_qt	qt 0.1	qt 0.2	qt 0.3	qt 0.4	qt 0.5	qt 0.6	qt 0.7	qt 0.8	qt 0.9	qt 1.0
Subscription										
0	0.931034	0.927806	0.908231	0.891718	0.885873	0.87992	0.873177	0.857523	0.833886	0.838752
1	0.068966	0.072194	0.091769	0.108282	0.114127	0.12008	0.126823	0.142477	0.166114	0.161248

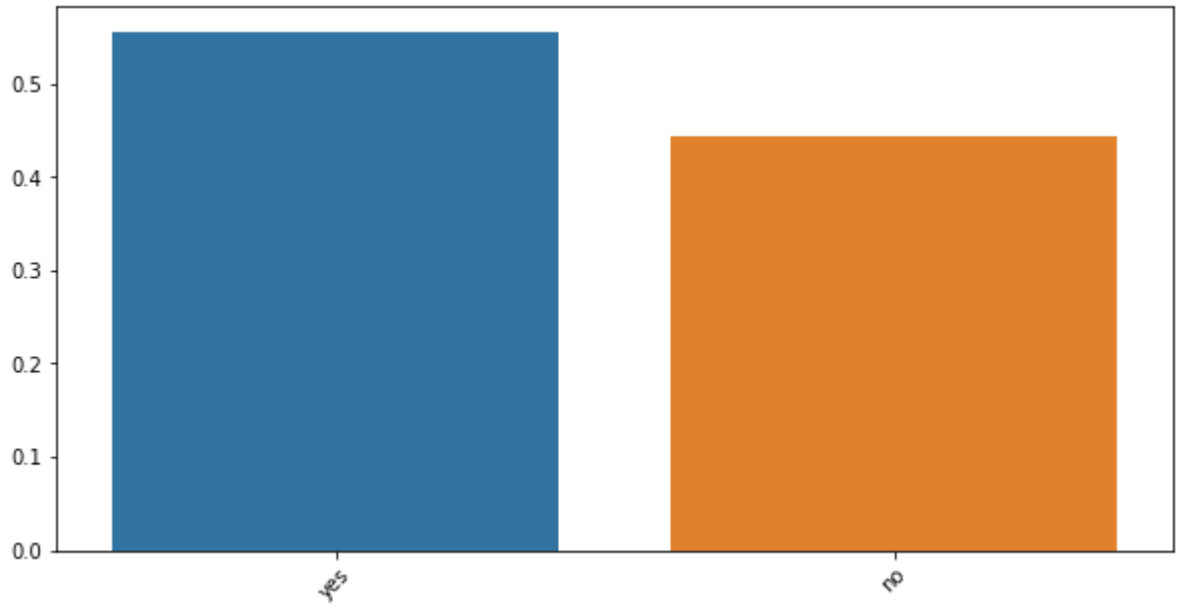
Balance_qt	qt 0.1	qt 0.2	qt 0.3	qt 0.4	qt 0.5	qt 0.6	qt 0.7	qt 0.8	qt 0.9	qt 1.0
Subscription										
0	6777	1645	4127	4027	3982	3979	3952	3870	3770	3792
1	502	128	417	489	513	543	574	643	751	729

La ventaja de la partición en cuantiles es que quedan grupos de igual tamaño. La desventaja es que los últimos cuantiles cubren un gran espacio de balances donde hay partes heterogéneas.



Otra opción es considerar el balance negativo o cero como una variable categórica y usar solo los balances positivos como variable numérica.

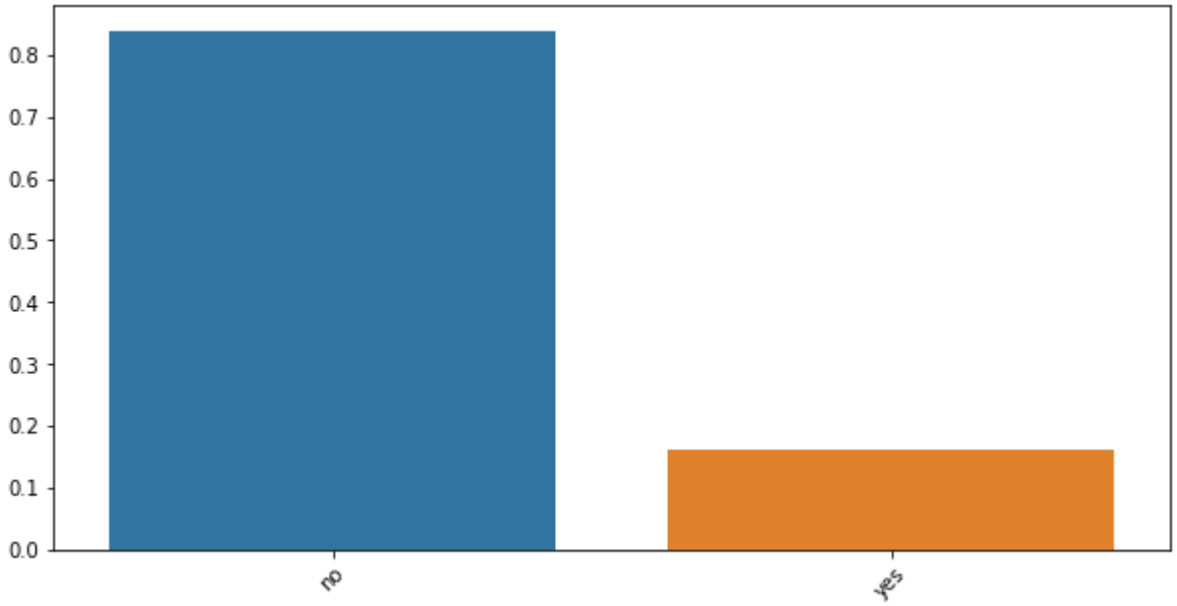
Housing Loan: Vemos una diferencia significativa entre las categorías de esta variable contra la variable respuesta.



Housing Loan	no	yes
Subscription		
0	16727	23195
1	3354	1935

Housing Loan	no	yes
Subscription		
0	0.832976	0.923
1	0.167024	0.077

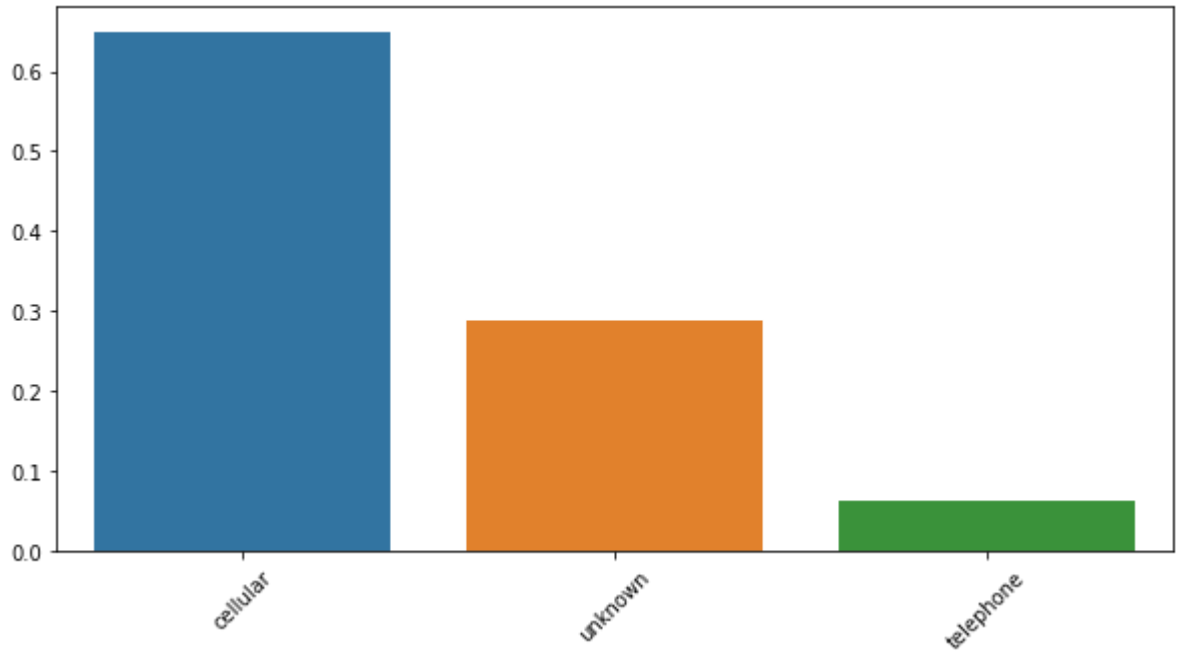
Personal Loan: Tiene una distribución similar a la anterior.



Personal Loan	no	yes
Subscription		
0	33162	6760
1	4805	484

Personal Loan	no	yes
Subscription		
0	0.873443	0.933186
1	0.126557	0.066814

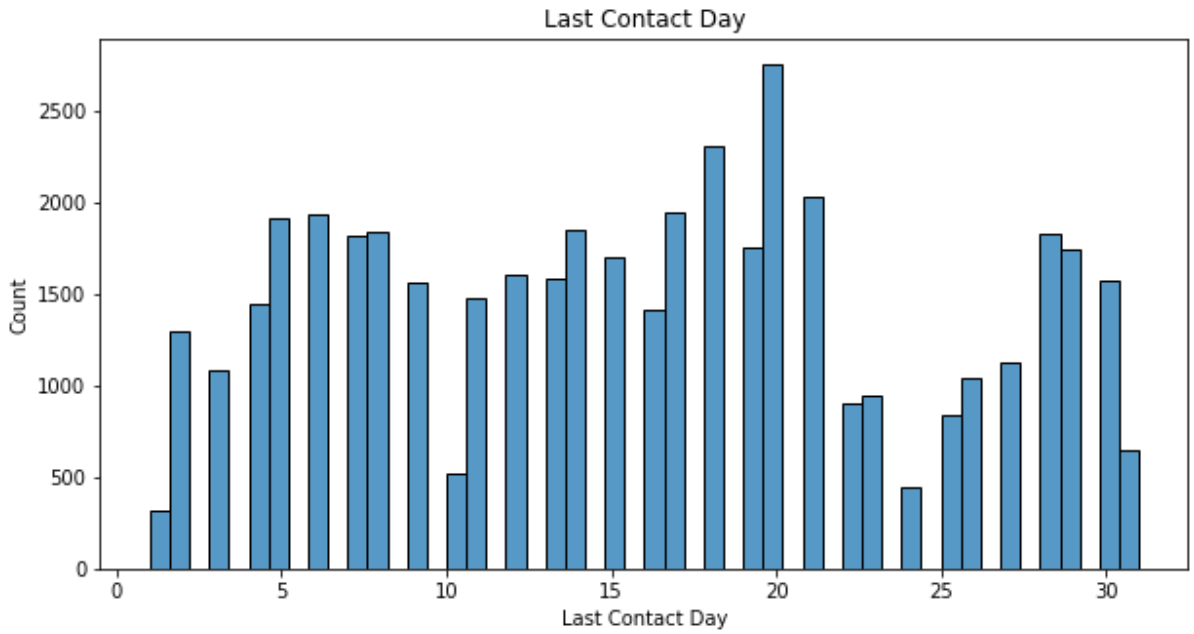
Contact: Hay una clara diferencia en la categoría “unknown” pero no está clara la interpretación de ese valor faltante. Vemos que todas las observaciones tienen duración de último contacto, por lo que hubo un contacto pero no se conoce la vía



Contact	cellular	telephone	unknown
Subscription			
0	24916	2516	12490
1	4369	390	530

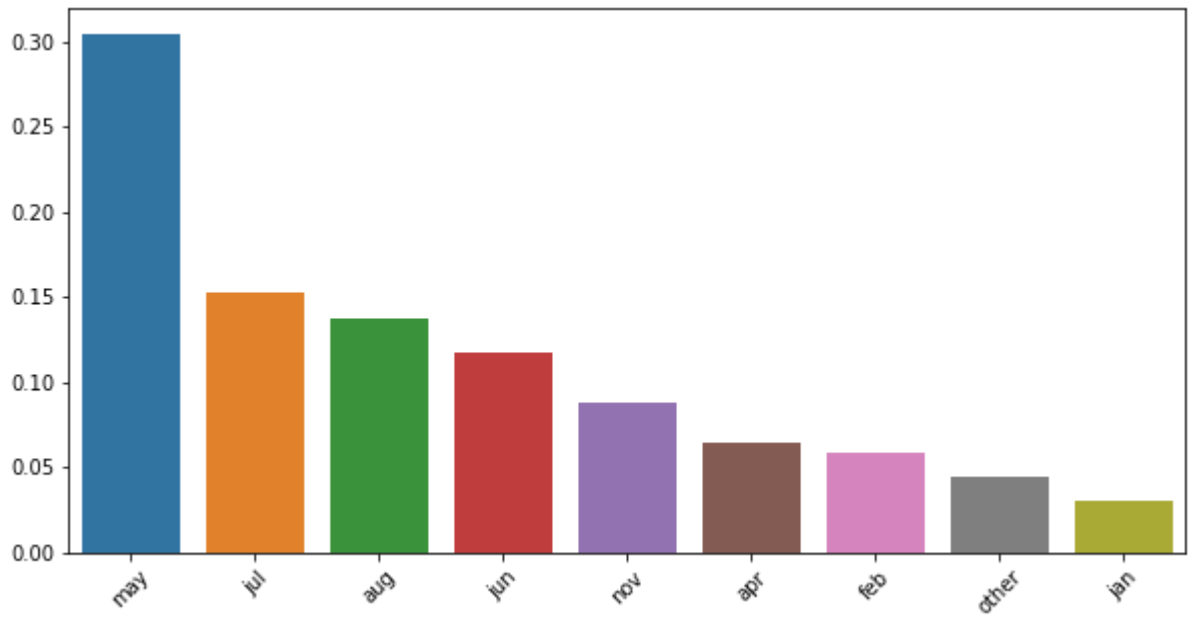
Contact	cellular	telephone	unknown
Subscription			
0	0.850811	0.865795	0.959293
1	0.149189	0.134205	0.040707

Last Contact Day: Es una variable numérica que no presenta una distribución interesante respecto a la respuesta. Pensamos en dividirla según la semana del mes. Se puede ver que en la tercera semana del mes, disminuye la proporción de gente que accede al plazo fijo.

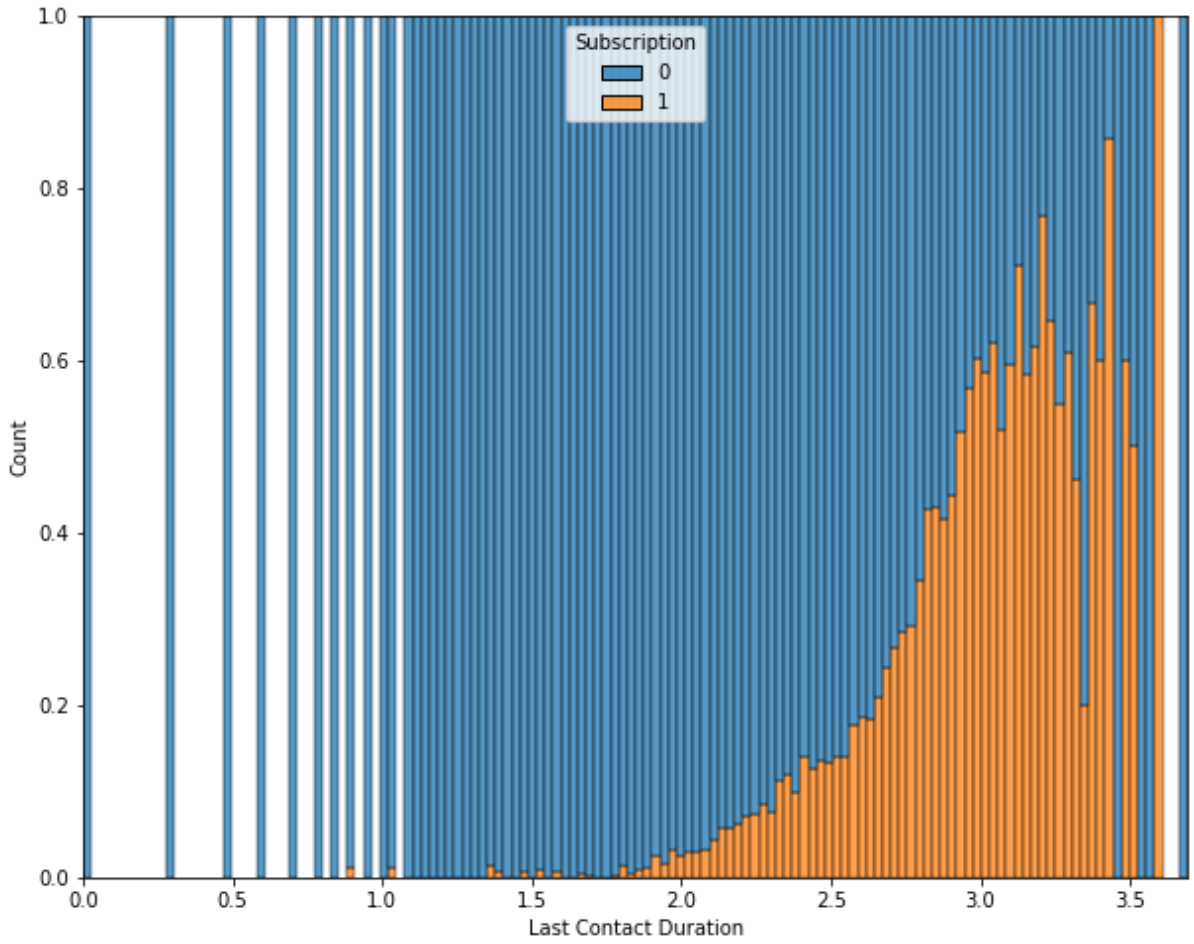


Last Contact Day	(0, 7]	(7, 14]	(14, 21]	(21, 31]
Subscription				
0	0.874158	0.868129	0.90295	0.879866
1	0.125842	0.131871	0.09705	0.120134

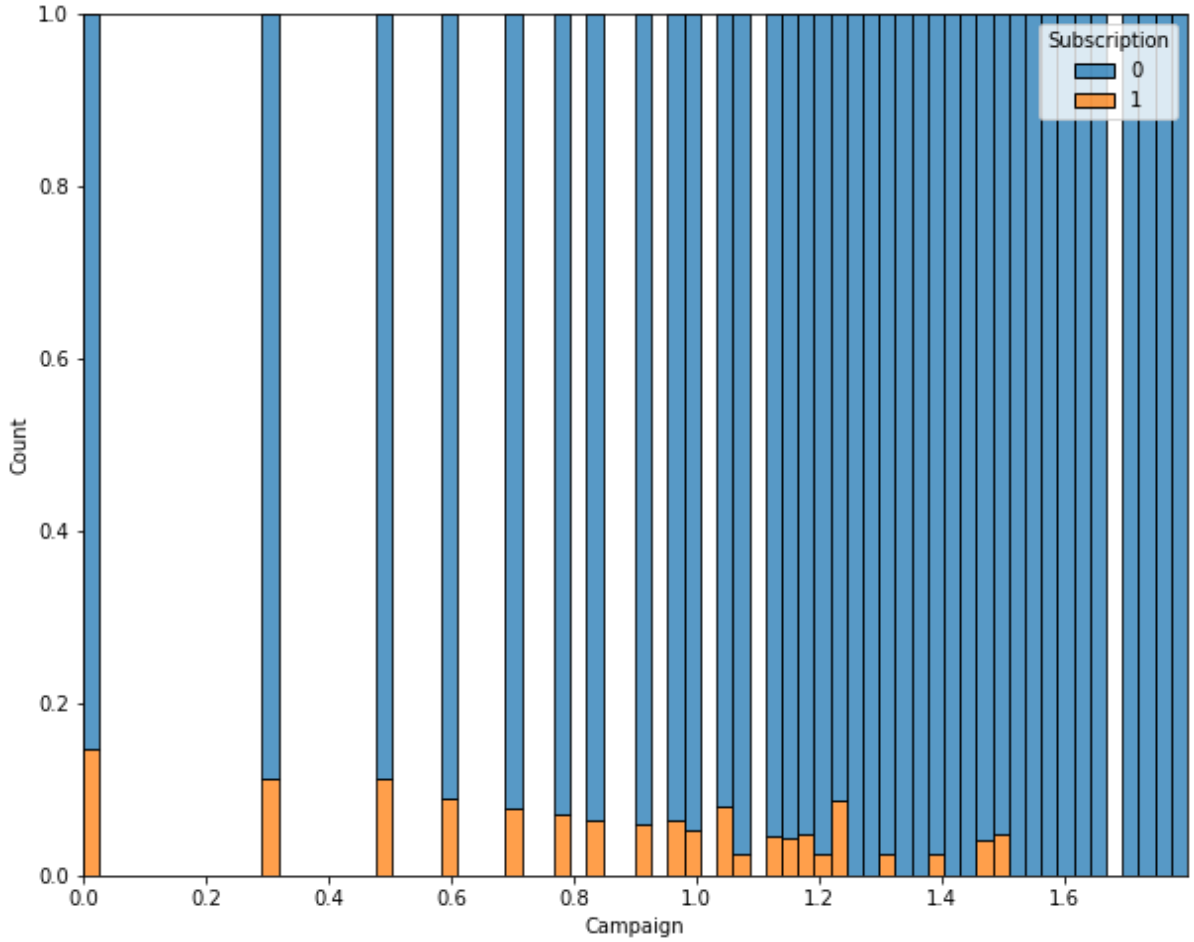
<u>Last Contact Month:</u> Los meses de menor frecuencia (octubre, septiembre, marzo, diciembre) tienen una distribución de la variable respuesta muy diferente. Los agrupamos en una categoría llamada “otros”



Last Contact Duration: Vemos un cambio en la distribución para los casos donde hubo suscripción. A mayor duración del último contacto, podemos ver un aumento en la proporción de éxito en la campaña.



Campaign: A mayor cantidad de contactos hay una menor proporción de éxito en la campaña.



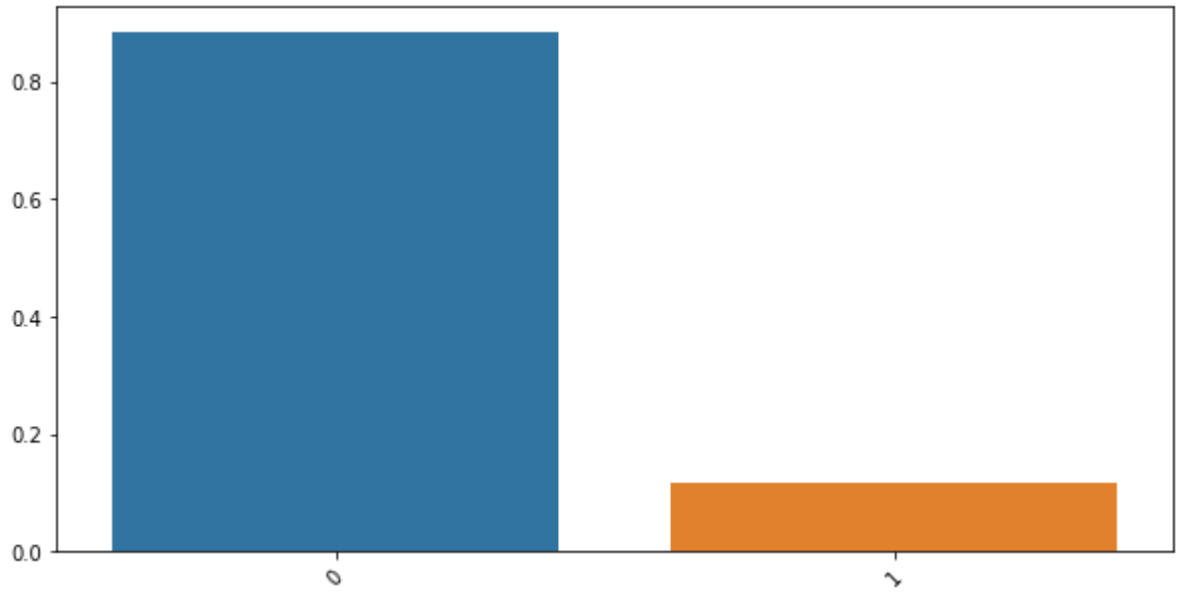
Variables P (campañas anteriores):

Estas variables tienen una alta cantidad de valores faltantes porque la gran mayoría de los clientes de la campaña actual no participaron de campañas anteriores. Es importante remarcar que tiene demasiados valores faltantes como para utilizarlas pero que aportan mucha información importante, por lo que más adelante mostraremos cómo incluirlas en el análisis.

Poutcome: esta variable resulta importante, dependiendo del resultado de la campaña anterior vemos que hay una variación en la proporción respecto a la variable respuesta.

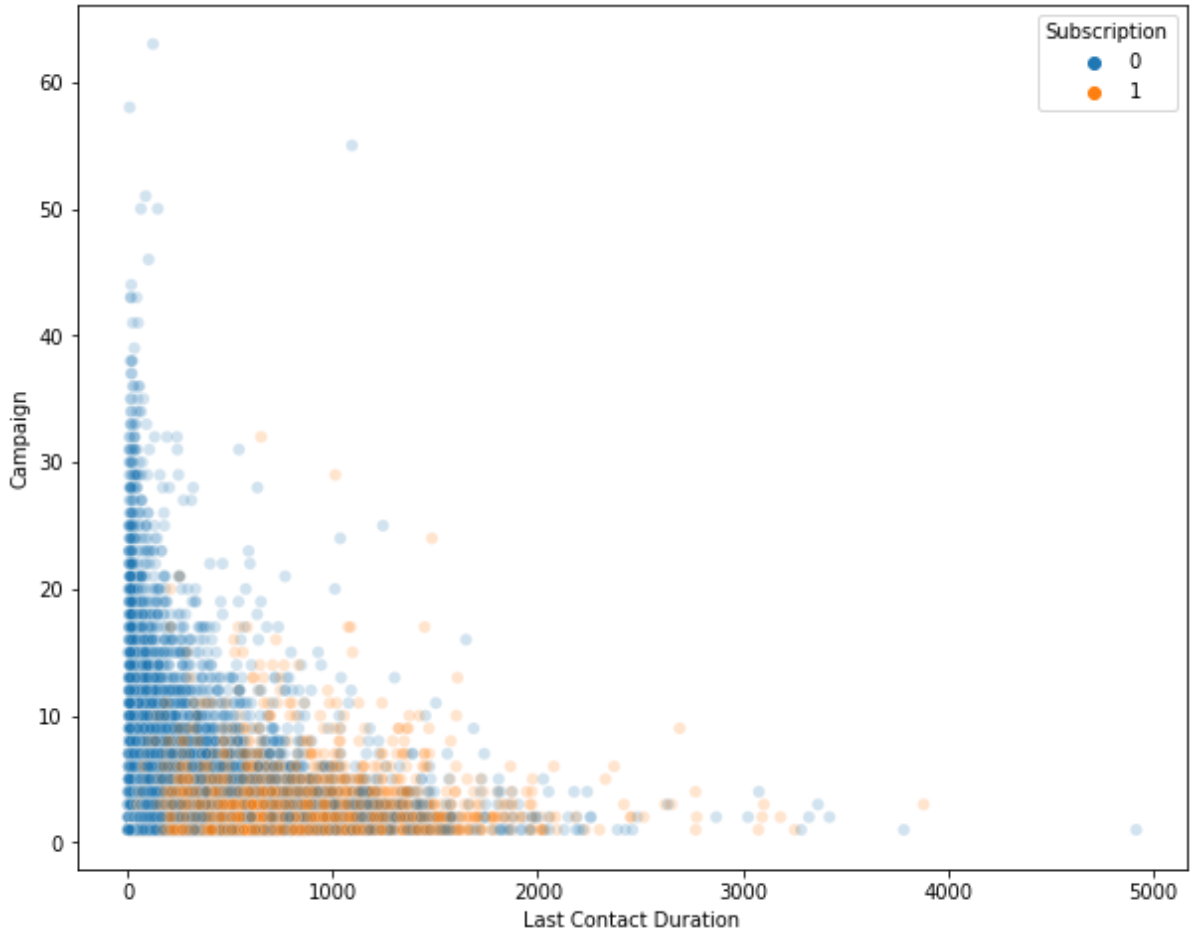
Poutcome	failure	other	success	unknown
Subscription				
0	0.873903	0.833152	0.352747	0.908385
1	0.126097	0.166848	0.647253	0.091615

Subscription (variable respuesta): tiene una distribución asimétrica, con solo el 10% de los casos positivos



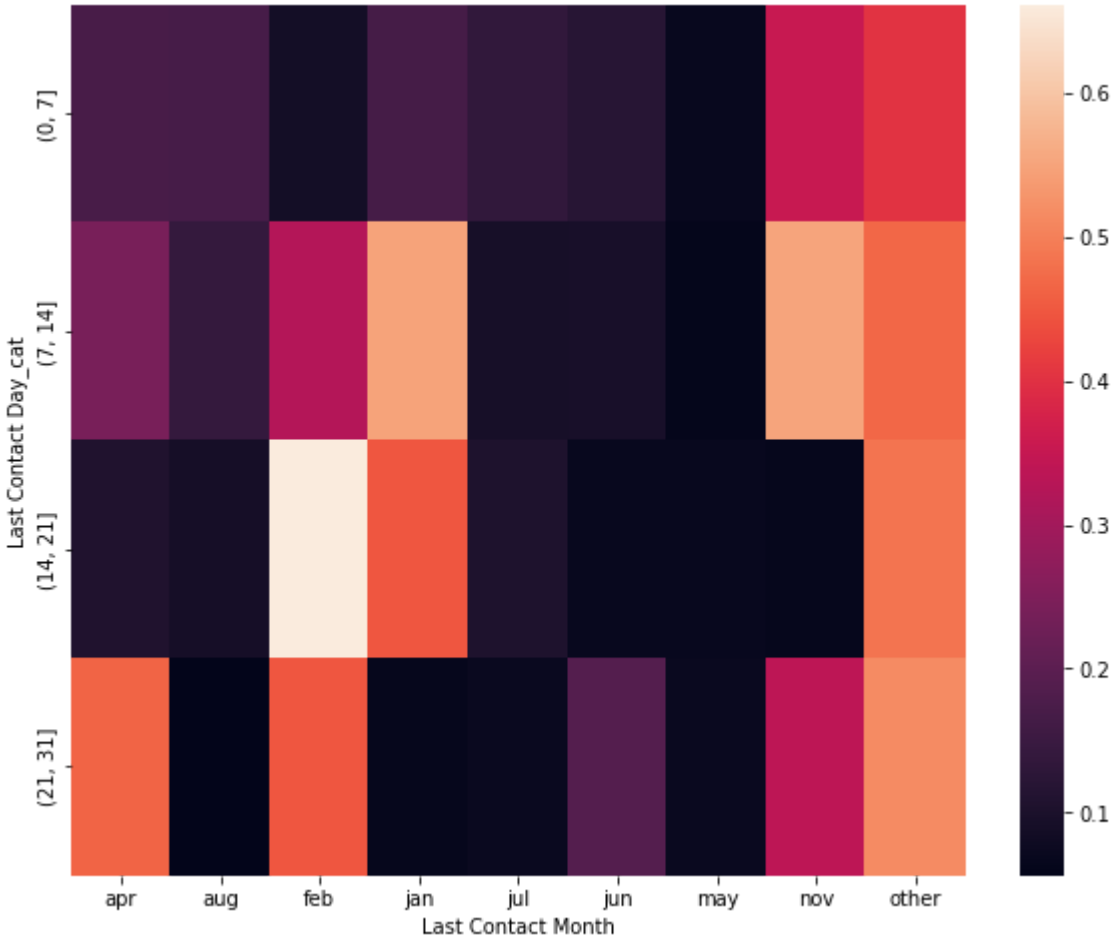
Otros gráficos:

- A mayor duración del último contacto y a menor número de contactos en la campaña, parece más probable el acceso al plazo fijo



- Heatmap para entender la relación de la proporción de la variable respuesta (color) según la semana y el mes del último contacto. Los meses entre septiembre y marzo

(varios agrupados en “other”) tienen una proporción más alta de éxito en la suscripción.



Modelos

Luego de analizar todas las variables del dataset realizamos las transformaciones que fuimos mencionando previamente, creamos las variables dummy a partir de las variables categóricas e imputamos algunos de los valores faltantes como los de la variable Education.

A partir de acá armamos los datasets de Train y Test para evaluar y seleccionar el mejor modelo predictor dejando afuera las variables P (campañas previas) por tener demasiados valores faltantes. Estos son nuestros datasets **base**.

Los modelos probados fueron:

- Regresión logística (penalización L2)
- Árbol de clasificación
- Random Forest
- Gradient Boosting

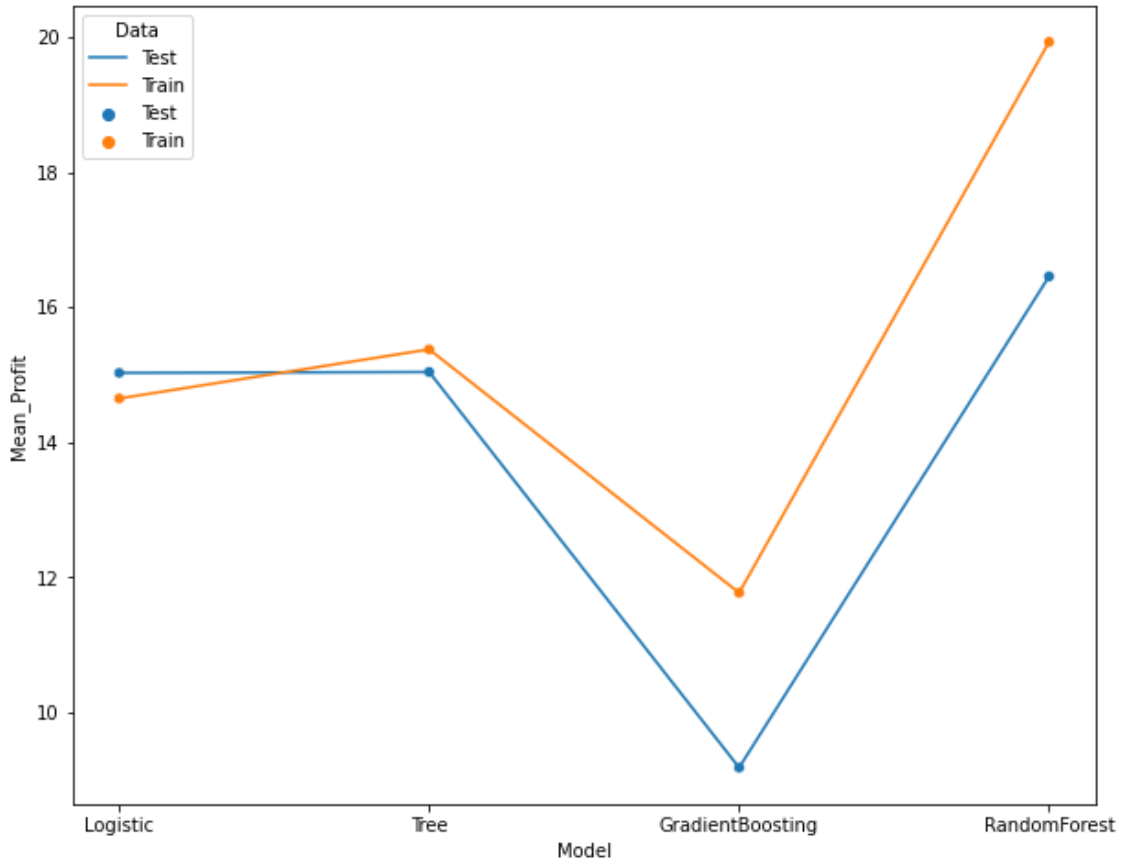
Para cada modelo se usó 5 fold Cross Validation con el train set optimizando la función de profit definida por la matriz de ingresos y costos del enunciado. De esta manera, la métrica optimizada fue el profit promedio para la selección de mejores hiperparametros.

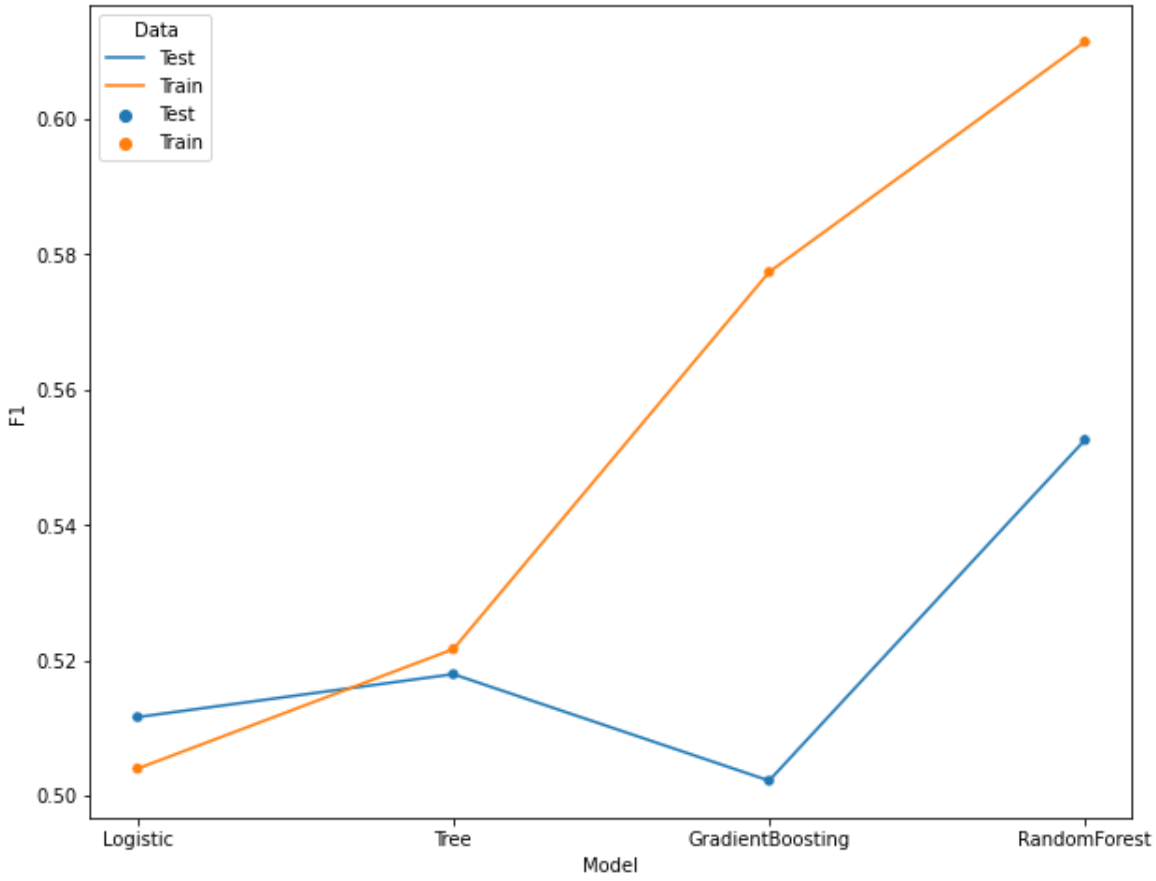
Siempre que estuviera disponible se optimizó el hiperparámetro class-weight para ponderar con un mayor peso a la clase 1 de la variable respuesta por el desbalance que tiene.

Obtenidos para cada modelo su mejor set de hiperparámetros, siendo estos los que daban el mejor profit promedio por validación cruzada, se entrenaron con el set de entrenamiento completo y se usaron para realizar predicción sobre el set de Test.

Podemos ver los resultados para el valor promedio de profit y la métrica f1 tanto para el set de entrenamiento como para el set de evaluación para cada modelo.

Podemos ver resultados parejos entre la Regresión Logística y el Árbol de Clasificación, un menor desempeño para el modelo Gradient Boosting y la mejor performance para el modelo Random Forest. Si bien este último posee una mayor diferencia entre el resultado de entrenamiento y de testeo el set de hiperparámetros elegido es el que mejor resulta de los 2466 evaluados.

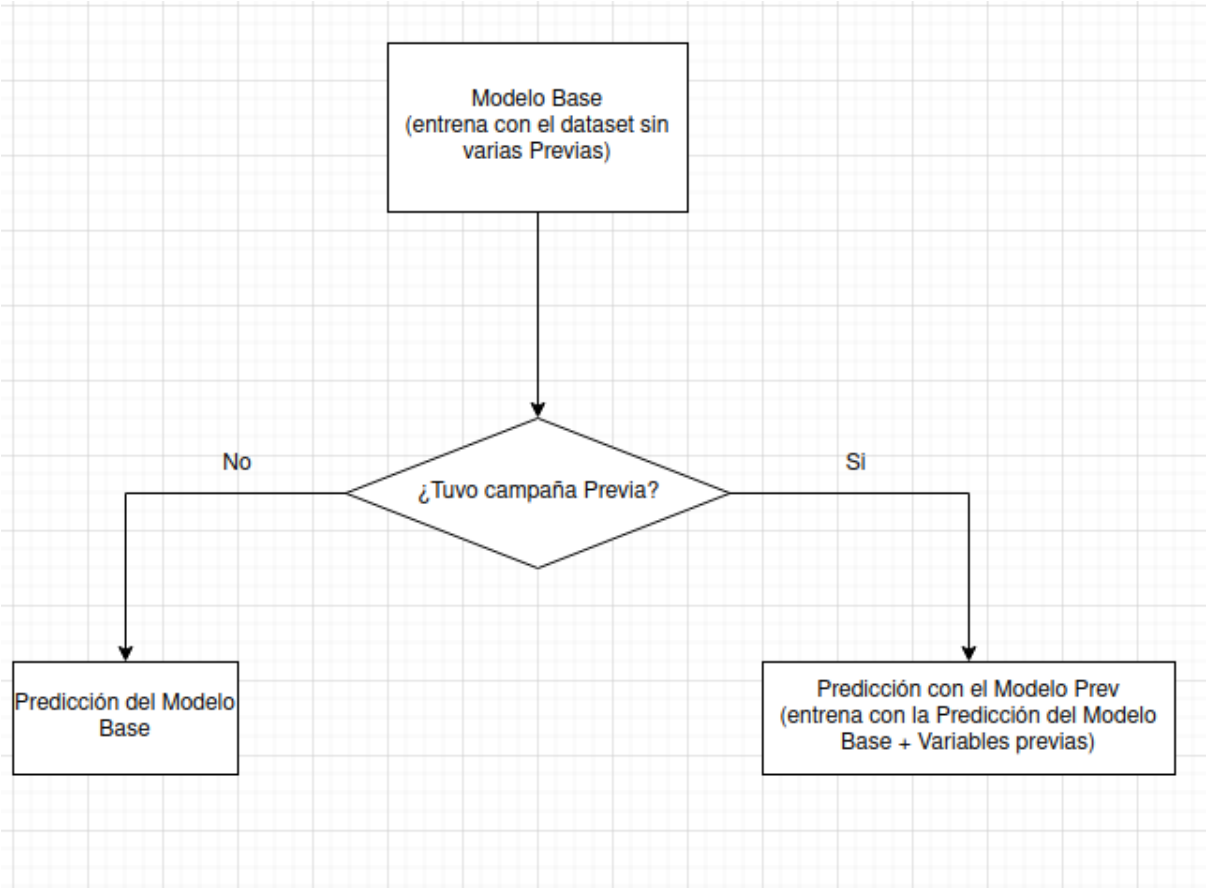




Modelo Cascada

Para poder incluir en nuestras predicciones las variables de campañas previas (variables P) pensamos en usar dos modelos en serie, de manera que el primero de los modelos sea el mejor de los Random Forest obtenidos y, si la observación posee la información de campañas previas, usar un nuevo modelo que ajuste la predicción del modelo base usando la información de dichas variables no incluídas originalmente.

Vemos un esquema de cómo quedaría la predicción usando este método. Pensamos que este método, a pesar de requerir más tiempo de entrenamiento y mayor cantidad de código podría traer una ventaja en las predicciones por estar usando información relevante de campañas anteriores.



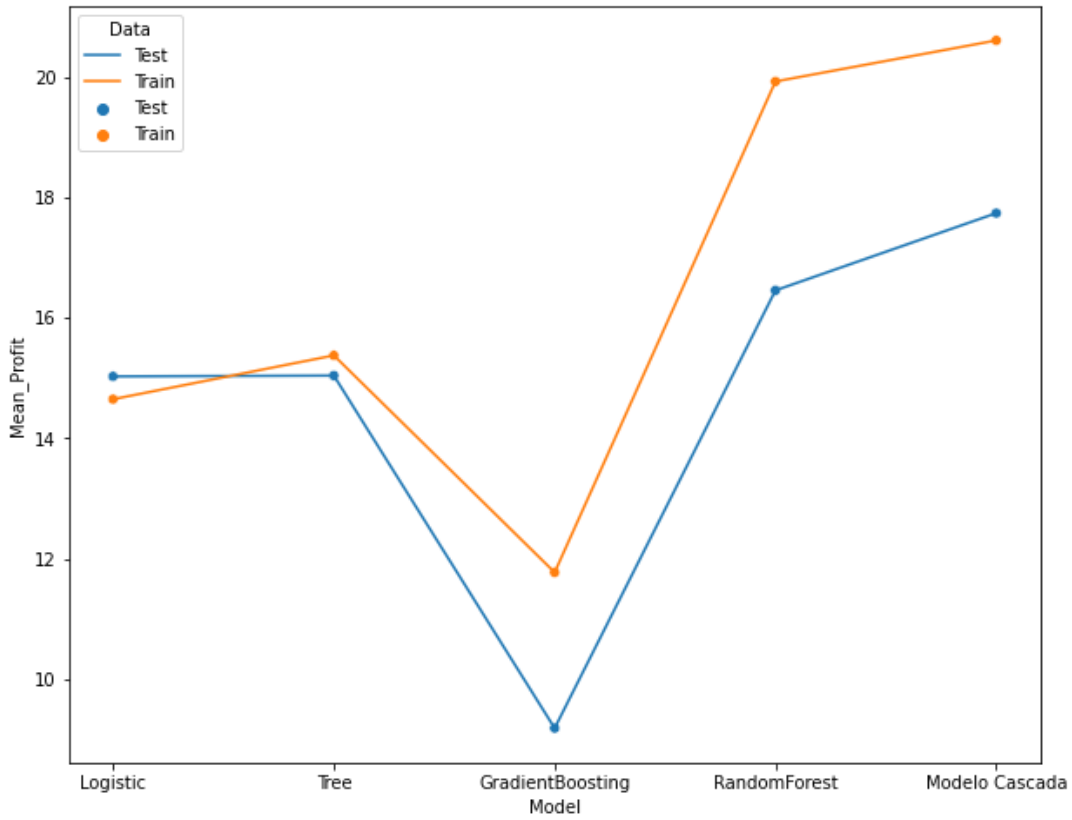
Para entrenar a el segundo modelo de ajuste o modelo **prev** creamos un dataset con las variables de campañas previas y la predicción del modelo base (mejor Random Forest), además de la variable respuesta.

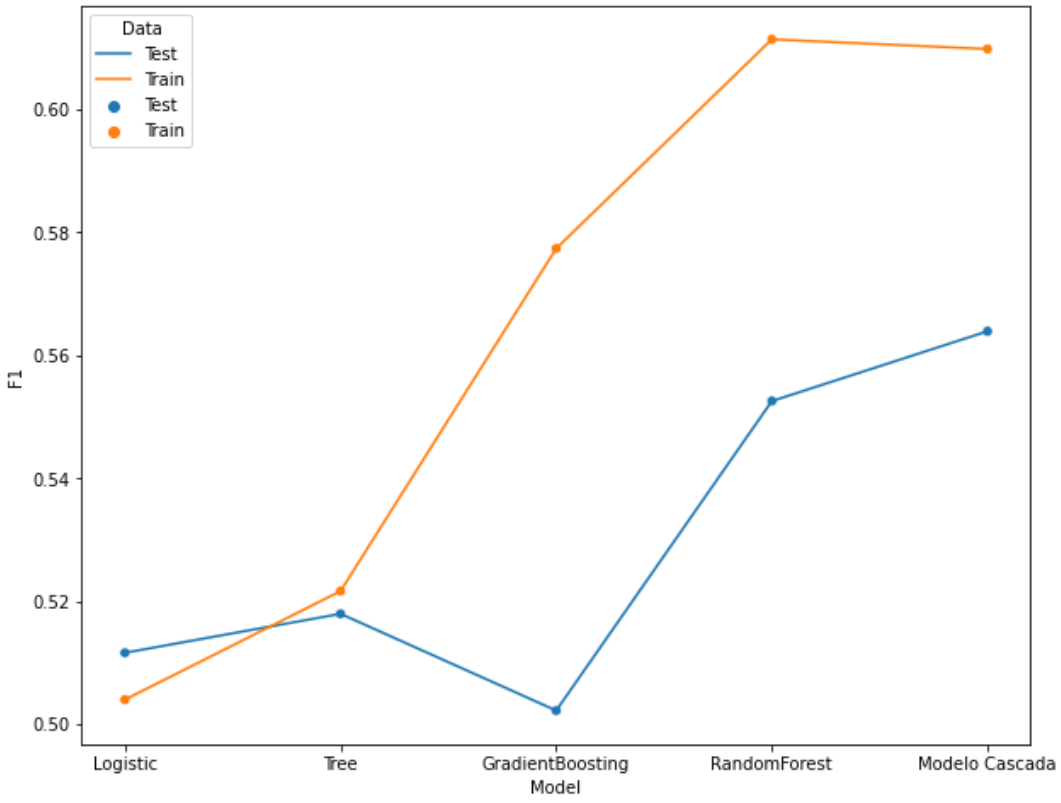
Con estos datos entrenamos un modelo de Random Forest al que llamaremos modelo prev, mientras que el modelo original lo consideramos el modelo base.

Creamos la función cascadePrediction que permite introducir un set de datos a predecir y devuelve la predicción con la condición de que si la observación no tiene información previa la predicción final será la del modelo base mientras que si tiene información de campañas anteriores se realiza la predicción del modelo base y esta se agrega al set de variables previas para realizar la predicción final con el modelo prev.

Los resultados muestran un incremento en las métricas de evaluación sobre datos no vistos por los modelos en el entrenamiento. Vemos una mejora en el profit medio por observación pasando de 16.46 euros a 17.74 euros y un valor mayor en la métrica f1.

Para la predicción de futuras observaciones nuevas entrenamos los dos modelos (tanto base como prev) con la totalidad de los datos dados para el Trabajo Práctico y creamos las funciones **prepareDataPrediction** que recibe un dataset con las mismas columnas que el entregado en el enunciado y devuelve el dataset transformado para el uso del modelo; y la función **cascadePredictionFull** que recibe el dataset transformado de variables explicativas y devuelve la predicción discreta o de probabilidad realizada por los modelos finales.


Data
Test
Train
Test
Train
Mean_Profit
20
18
16
14
12
10
Logistic
Tree
GradientBoosting
RandomForest
Modelo Cascada
Model


Data
Test
Train
Test
Train
F1
0.60
0.58
0.56
0.54
0.52
0.50
Logistic
Tree
GradientBoosting
RandomForest
Modelo Cascada
Model