Presentación del Proyecto Final

Análisis Integral de Generación de Default

Autor: Angel Lucero

ÍNDICE

OBJETIVO

El Proyecto Final de Data Science busca crear un algoritmo de clasificación de Machine Learning, para aumentar la precisión en un 15% de sí un usuario es un potencial deudor o no dentro de los próximos siete meses.
Con el objetivo de determinar si es viable que reciba un nuevo crédito o aumentar el que ya tiene.

01

CONTEXTO Y **AUDIENCIA**

Contexto Comercial: Nuestro cliente es un banco y está preocupado porque quiere analizar si sus clientes son propensos a hacer default o no en el próximo mes.

Problema Comercial: Hay diferentes factores que influyen para que un cliente pueda tener más límite de crédito.

Contexto Analitico: Hacer un análisis más profundo con este tipo de características, y podremos dar recomendaciones al banco dado el resultado.

Audiencia: Este análisis va dirigido a la parte gerencial que se encarga de tomar las decisiones.

02

METADATA

En total el DataSet tiene unos 30.000 registros y 24 columnas:

- Los registros se componen por todos los clientes.
- Una columna contiene el límite de crédito de cada cliente.
- 4 de las columnas poseen datos demográficos del cliente.
- 6 columnas se conforman por el Historial de Pagos Pasados de los últimos 6 meses.
- 6 columnas se conforman por el Importe del estado de cuenta de los últimos 6 meses.
- 6 columnas se conforman por el Importe del pago anterior de los últimos 6 meses.
- Una columna posee los datos de que si el cliente hizo default o no.

03

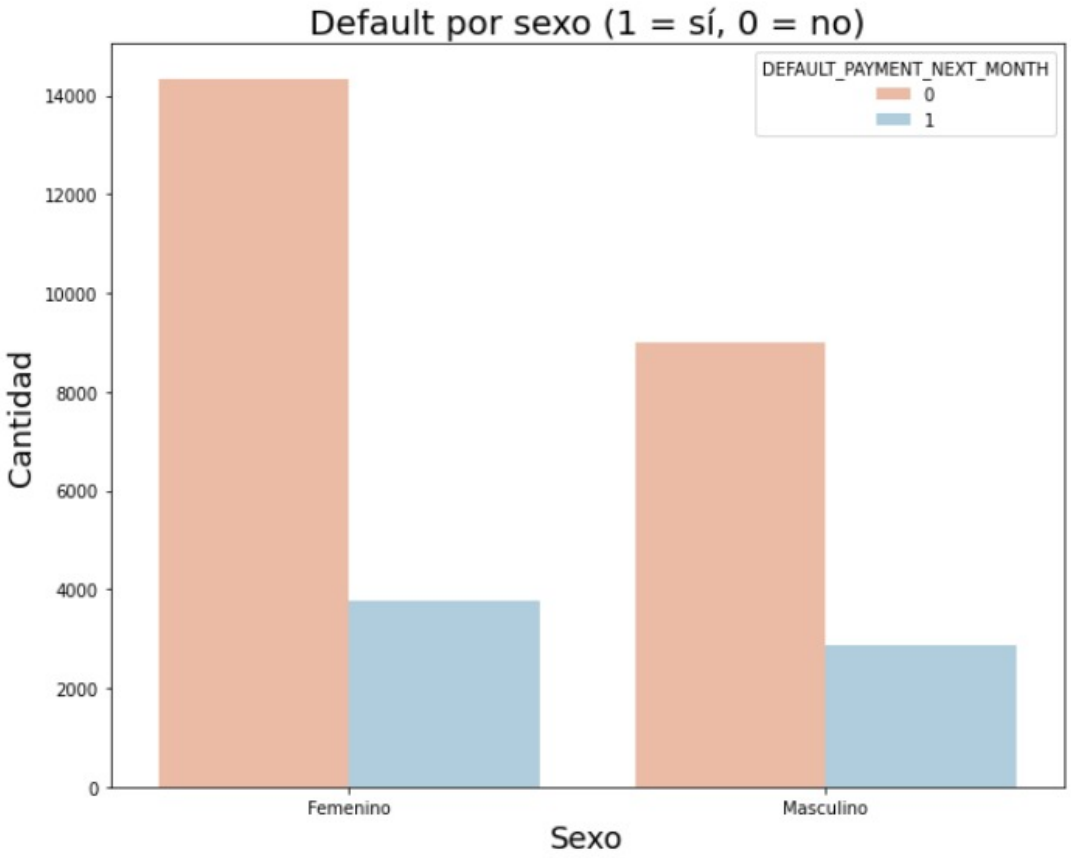
PREGUNTAS DE **INTERÉS**

Preguntas generadoras de insights:

- **¿Qué género es el que más influye a la hora de hacer default?. También segmentarlo dentro del nivel educativo y estado civil.**

INSIGHT 1

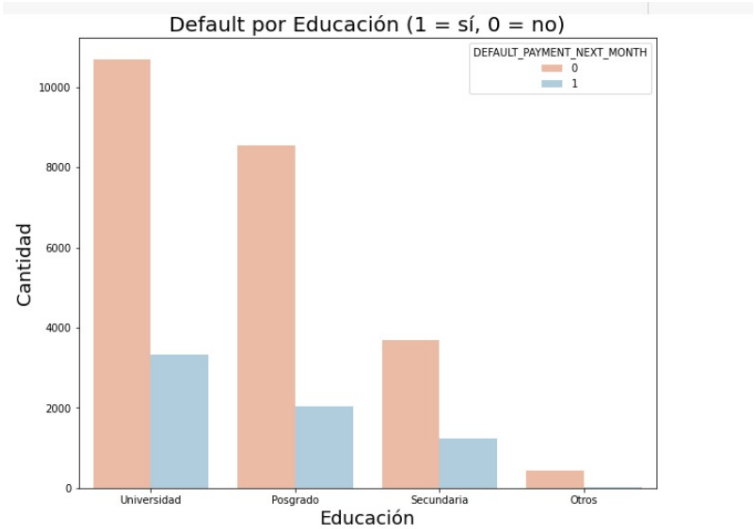
¿Qué género influye más en el Default?

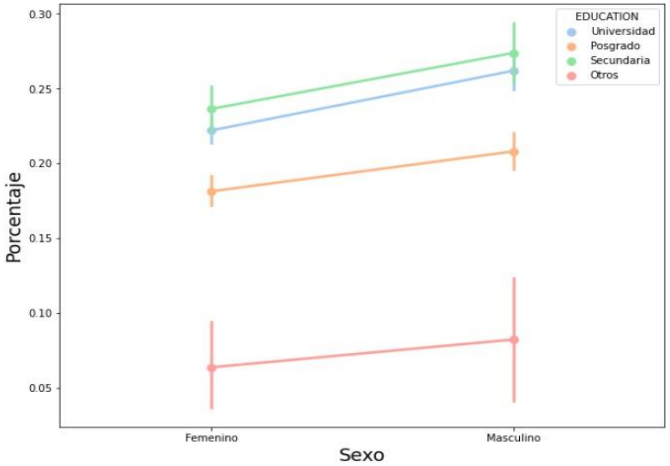


Podemos ver que el género **Masculino** tiende a tener mayor, su default es de un **24%** y el del género **Femenino** un **21%**.

Lo que sí hay que tener en cuenta es que hay mayor cantidad de género femenino que masculino, por lo que teniendo una muestra mayor, el género fémenino genera menos default al siguiente mes.

En educación, ¿Quién tiende a generar default?



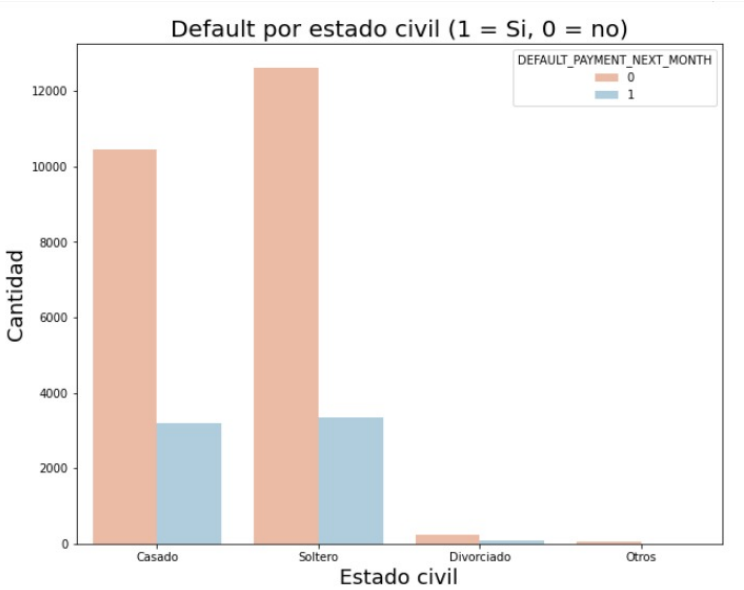


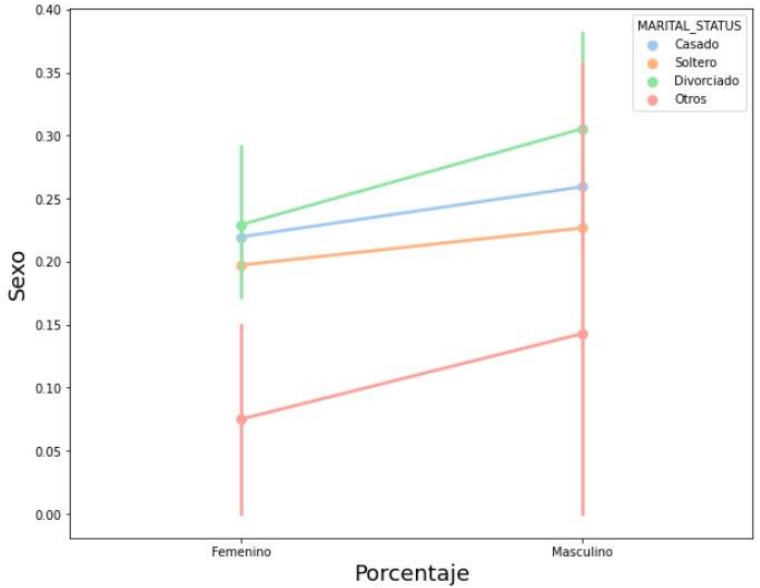
Vemos que según la variable de Nivel Educativo de los clientes, las personas que tienen **Nivel de Secundaria**, tienen una **Mayor** probabilidad de generar default, respecto al resto de los niveles educativos.

Haciendo una apertura según el sexo, vemos que hay una diferencia de 2 puntos porcentuales, un 24% para las mujeres y un 26% para el hombre.

Pero si vemos la categoría de **Universidad**, tiene mayor cantidad de personas que la categoría secundaria y no es tan lejano los porcentajes que maneja entre uno y el otro sobre los que hacen default, ya que cuenta con un 22% para las mujeres y un 24% para los hombres.

En estado civil, ¿Quién tiende a generar default?





Si bien la información que tenemos al respecto de los diferentes estados civiles no es proporcional, podemos ver que los clientes que son **Solteros**, tienen una menor participación en el default generado en el próximo mes, que las personas **casadas.**

También, en la categoría de **Divorciados** hay mayor porcentaje proporcional de default, un 25% para género femenino y un 30% para género masculino.

RECOMENDACIONES

- **En la educación, vemos que los que tienen titulación secundaria tienden a tener mayor default. Podemos recomendar ver la cantidad de personas que tienen titulación con secundaria y tratar de darle menos límite de crédito. También para generar mayor impacto en bajar el default se puede recomendar concentrarse en las personas con titulación universitaria y de posgrado.**

- **En el estado civil, se recomienda seguir limitando el límite de crédito para las personas divorciadas, ya que son las que generan mayor default.**

05

Implementación de modelos de Machine Learning

Para abordar nuestro problema comercial de que clientes son propensos a hacer default el mes próximo en su resumen de tarjeta de crédito, implementamos los siguientes modelos de clasificación supervisada con técnicas de cross validation e hyperparameter tunning,

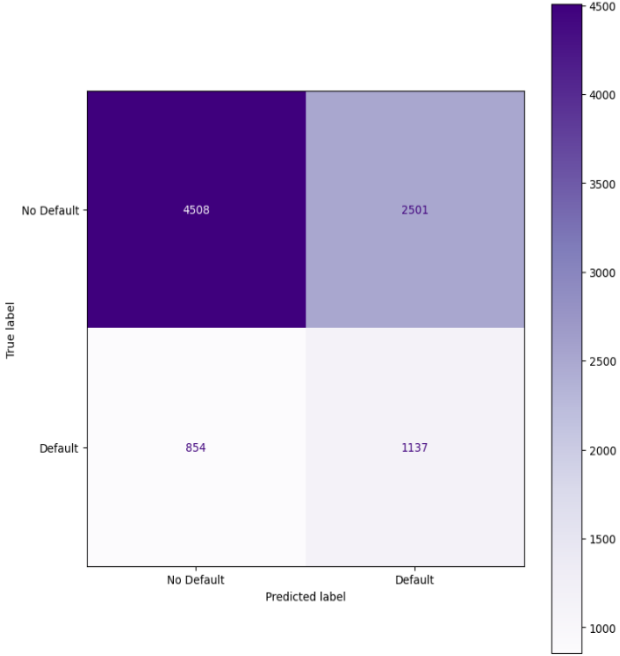
Nuestro objetivo es obtener un recall mínimo de 85%, ya que; en el área de las finanzas decimos que la sensibilidad(recall) es la capacidad de poder detectar correctamente si clientes son propensos a hacer default el mes próximo.

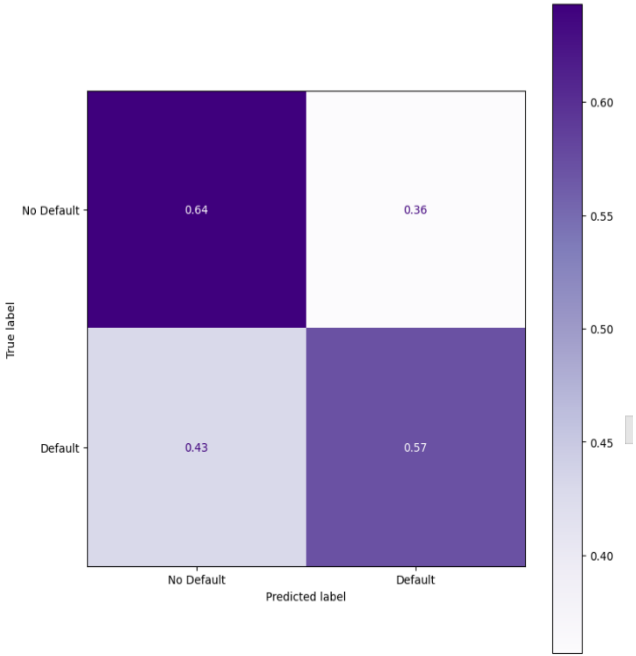
- **Decision Trees**
- **Random Forest**
- **Light GBM**
- **RandomizedSearchCV y GridSearchCV**

Modelo 1: Decision Tree

En el primer modelo podemos ver que el 43% de los casos predijo que los clientes no hacen default y si hacen. (Falsos Negativos).

El 57% de los casos que predijo que los clientes hacen default, los predijo correctamente; hacen default.(Verdaderos Positivos)

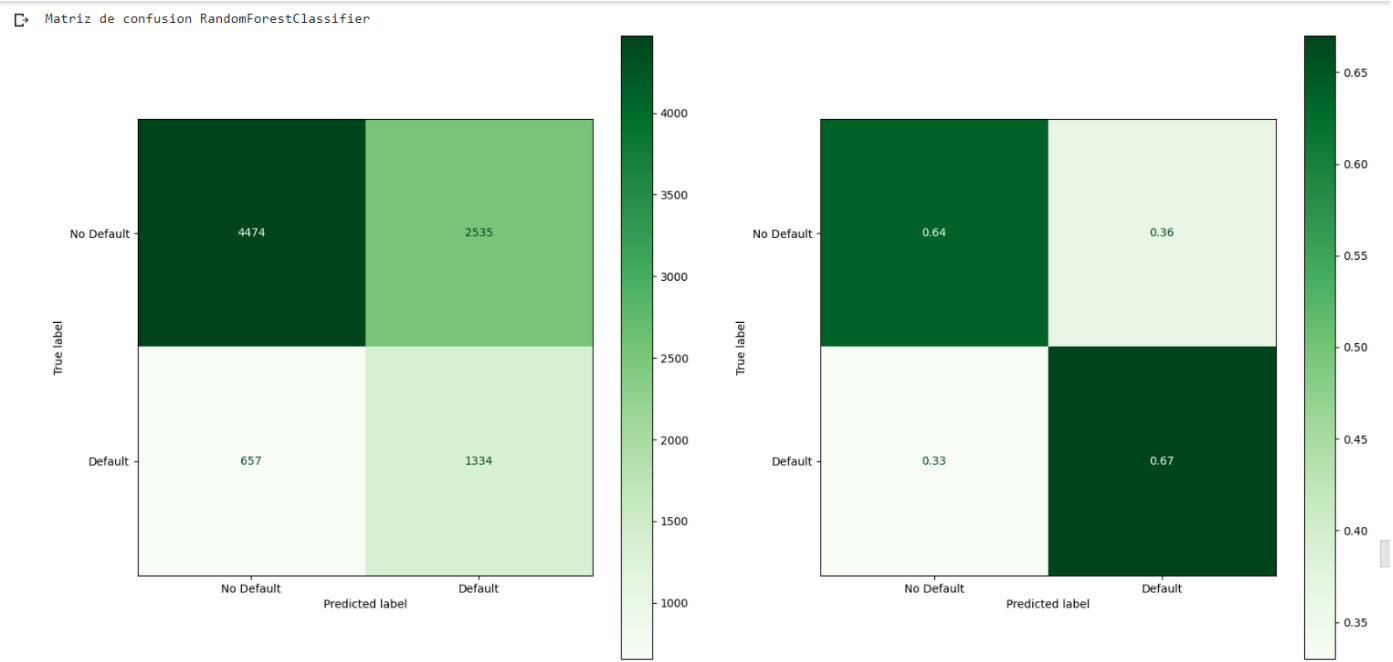




Modelo 2: Random Forest

En el segundo modelo podemos ver una disminución en los (falsos negativos) en comparación al modelo anterior, donde el 33% de los casos predijo que los clientes no hacen default y si hacen.

También aumentó en los (verdaderos positivos) en el que el 67% de los casos que predijo que los clientes hacen default, los predijo correctamente; hacen default.

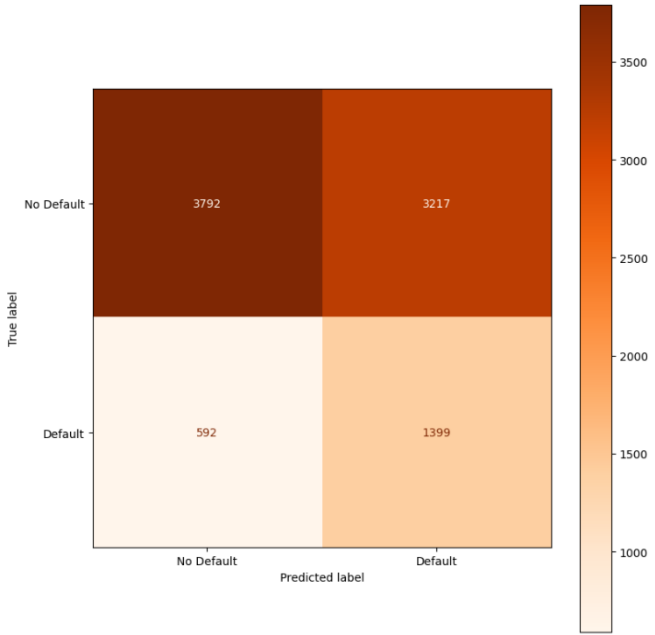


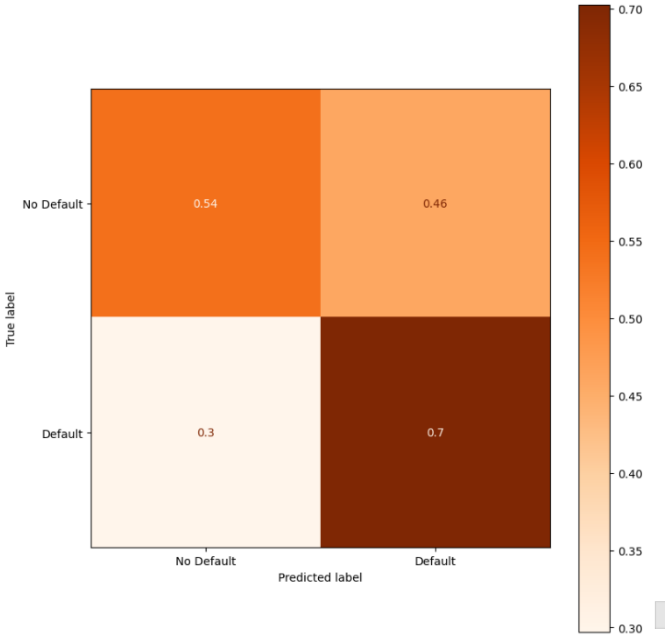
Modelo 3: LightGBM

En el tercer modelo podemos ver que disminuyó aún más en los (falsos negativos) en comparación a los modelos anteriores, donde el 30% de los casos predijo que los clientes no hacen default y si hacen.

También aumentó en los (verdaderos positivos) en el que el 70% de los casos que predijo que los clientes hacen default, los predijo correctamente; hacen default.

Matriz de confusion LGBMClassifier



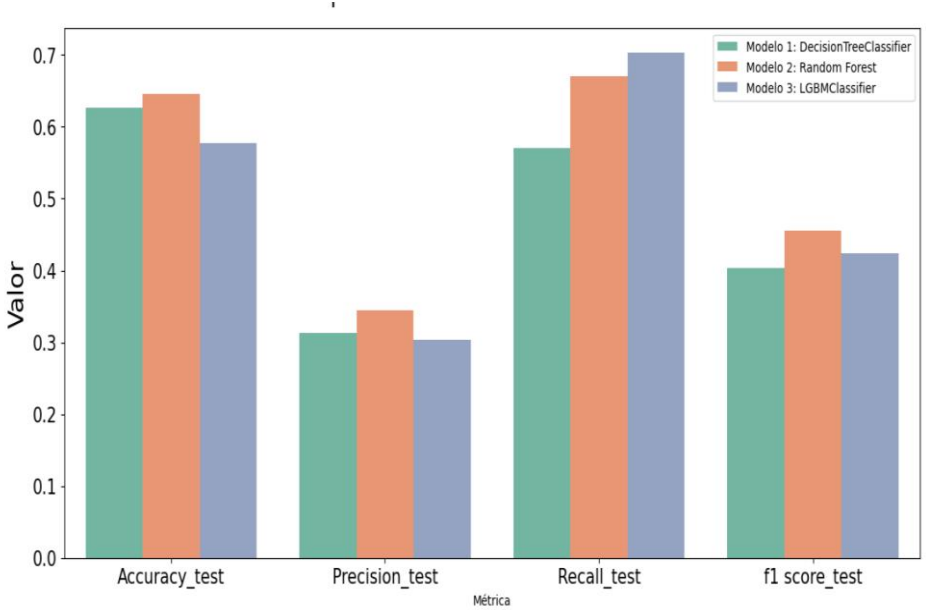


06 Comparamos las métricas obtenidas de los modelos

Basado en los resultados de los modelos y las métricas obtenidas, podemos concluir lo siguiente para el problema de clasificación binaria de predecir si un cliente hará default o no el mes próximo:

El desempeño de los modelos evaluados varía en precisión, recall y puntuación F1 en el conjunto de prueba. El Modelo 2, Random Forest, muestra el mejor desempeño general con una precisión de 0.34, recall de 0.67 y puntuación F1 de 0.45.

	Modelo	Métrica	Valor
0	Modelo 1: DecisionTreeClassifier	Accuracy_test	0.627222
1	Modelo 2: Random Forest	Accuracy_test	0.645333
2	Modelo 3: LGBMClassifier	Accuracy_test	0.576778
3	Modelo 1: DecisionTreeClassifier	Precision_test	0.312534
4	Modelo 2: Random Forest	Precision_test	0.344792
5	Modelo 3: LGBMClassifier	Precision_test	0.303076
6	Modelo 1: DecisionTreeClassifier	Recall_test	0.571070
7	Modelo 2: Random Forest	Recall_test	0.670015
8	Modelo 3: LGBMClassifier	Recall_test	0.702662
9	Modelo 1: DecisionTreeClassifier	f1 score_test	0.403979
10	Modelo 2: Random Forest	f1 score_test	0.455290
11	Modelo 3: LGBMClassifier	f1 score_test	0.423490



07

Conclusión

Podemos concluir que aunque los modelos evaluados no alcanzaron el umbral mínimo de recall del 85%, se observó un progreso significativo en la identificación de casos de default. El Modelo 3, LGBMClassifier, mostró un recall más alto de 0.70, lo que indica una mejora notable en la identificación de casos de default en comparación con los otros modelos. Sin embargo, aún no llega al umbral deseado.

Por lo tanto, en base a los resultados obtenidos, estos modelos aún no ofrecen una predicción satisfactoria para determinar si un cliente tiene una alta probabilidad de caer en default el próximo mes.

Próximos pasos:

- Se sugiere seguir ajustando los hiperparámetros y explorar técnicas de selección de características para mejorar aún más el rendimiento del modelo.
- La obtención de más datos de la clase minoritaria.
- La exploración de técnicas de balanceo de clases podrían ser opciones adicionales para abordar el desequilibrio en futuras iteraciones.