Resumen de los Resultados

Grupo: Desirée Vera, Felipe Gómez, Harmynn Garrido, Diego Granados.

**Parte A: Modelos**

Se entrenaron dos modelos de clasificación para determinar los defaults. El primero consiste en modelo logit, en el cual se utilizaron todas las variables explicativas para predecir a la variable objetivo Default.

El segundo se basó en un árbol de decisión optimizado con un max depth = 7, min\_samples\_leaf = 10 , min\_samples\_split = 2 y random\_state = 21.

Como se mencionó en el Análisis Exploratorio de Datos, la variable ‘Default’ presenta una leve asimetría, dado que el 63% de los registros corresponde a 1 (hubo default). El desbalanceo sesga al indicador Accuracy hacia el valor mayoritario, por lo que este indicador debe ser contrastado con lo que diga el Recall y la Precision. Esto nos llevó a optimizar los modelos usando el F1 Score, previniendo cualquier sesgo que pueda producir el desbalanceo.

**Parte B: Resultados y escogencia de modelo**

Comparación de modelos:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Modelo | Muestra | AUC | KS | Threshold | Accuracy | Precision | Recall | F1 |
| 0 | logit\_sk | Train | 0.844878 | 0.513741 | 0.260706 | 0.763788 | 0.749782 | 0.940822 | 0.834508 |
| 1 | logit\_sk | Test | **0.838633** | 0.516637 | 0.239813 | **0.744807** | **0.730541** | 0.944966 | **0.824033** |
| 2 | tree | Train | 0.817260 | 0.488966 | 0.414634 | 0.754422 | 0.742369 | 0.937352 | 0.828544 |
| 3 | tree | Test | 0.810299 | 0.478983 | 0.403785 | 0.733747 | 0.717817 | **0.953925** | 0.819198 |

Como se aprecia, para la muestra Test, el modelo logit presenta mejor Accuracy que el modelo tree. Sin embargo, el Recall favorece al modelo tree y la Precision al modelo logit. En ese sentido, el F1 Score, que muestra un balance entre ambos termina favoreciendo al Modelo logit, que es el que escogimos.