

Distribución Variable Objetivo

La variable objetivo (y) presenta un desbalance de clases . El 89.05% de los registros pertenecen a la clase NO, mientras que el 10.95% restante corresponden a la clase Sí. Esta desproporción indica que la gran mayoría de los clientes no se suscribieron al plazo fijo. Este fuerte desbalance debe ser considerado cuidadosamente en el entrenamiento, ya que puede llevar a que el modelo sesgue sus predicciones hacia la clase mayoritaria (NO), resultando en un bajo Recall para la clase minoritaria (Sí), como se observó previamente.

Distribución Variables Predictoras Clave

1. Age

La distribución de la edad se concentra en la edad adulta temprana y media. El promedio de edad es de 40 años y la mediana es de 38 años, sugiriendo una distribución ligeramente sesgada a la derecha (ya que la media es mayor que la mediana). La moda es 32 años, lo que indica que la edad más frecuente en la base de datos es relativamente joven. El rango de edad es amplio, desde un mínimo de 8 años hasta un máximo de 88 años. La baja moda combinada con la media de 40 sugiere que la mayoría de los clientes son más jóvenes, aunque hay una cola de clientes mayores que eleva la media.

2. Duration

La variable muestra una distribución altamente concentrada y fuertemente sesgada a la derecha. La media es de 256 segundos, pero la mediana es de solo 181 segundos, y la moda es 77 segundos. El hecho de que la media sea significativamente más alta que la mediana y la moda indica la presencia de valores atípicos muy grandes que arrastran el promedio. Estos valores atípicos se confirman por el máximo de 3643 segundos.

Tabla de Métricas de Test and Score

Model	AUC	CA	F1
Logistic Regression	0.751	0.901	0.876
Naive Bayes	0.731	0.781	0.812
Random Forest	0.748	0.894	0.874

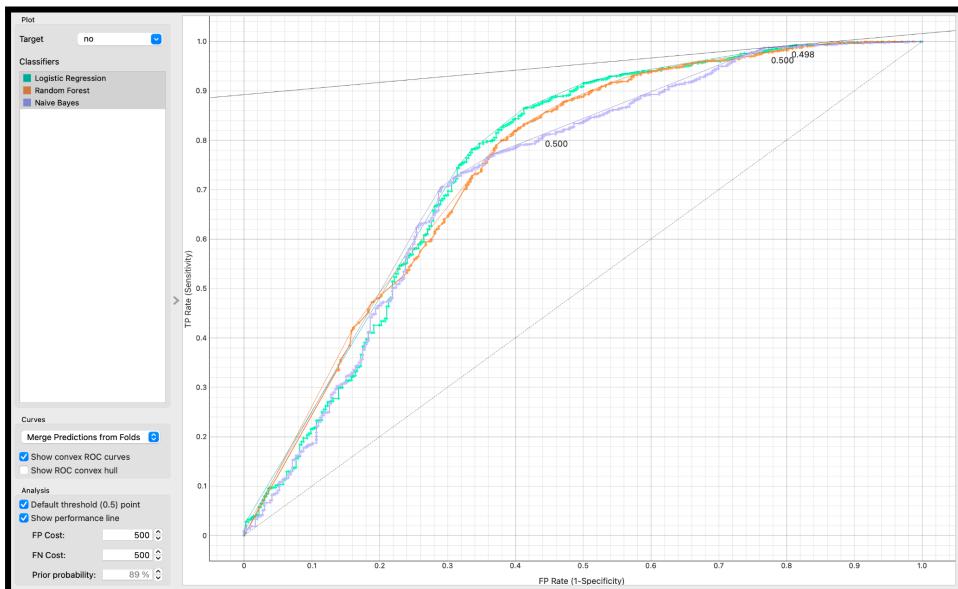
- **Modelo con el mejor AUC:** El modelo de **Regresión Logística** tiene el mejor **Área Bajo la Curva (AUC)** con un valor de **0.751**. Un AUC más alto indica una mejor capacidad del modelo para distinguir entre las clases positiva y negativa.
- **Modelo con el mejor F1-score:** El modelo de **Regresión Logística** también tiene el mejor **F1-score** con **0.876**. El F1-score es una media armónica de la Precisión (Precision) y la Exhaustividad (Recall), por lo que un valor más alto sugiere un mejor equilibrio entre ambas.

Matriz de Confusión (Regresión Logística)

		Predicted		Σ
		no	yes	
Actual	no	2898	32	2930
	yes	294	72	366
Σ		3192	104	3296

- **Verdaderos Positivos (TP):** 72
- **Verdaderos Negativos (TN):** 2898
- **Falsos Positivos (FP):** 32
- **Falsos Negativos (FN):** 294

Análisis de la Curva ROC



1. **Naive Bayes (color púrpura/violeta):** Es la peor curva, ya que está más cerca de la línea diagonal y se encuentra por debajo de las otras en gran parte del recorrido. Esto se confirma con su bajo AUC (0.731).
2. **Logistic Regression (color verde) y Random Forest (color naranja):** Ambas curvas son muy similares y se encuentran muy cerca de la esquina superior izquierda. Sin embargo, la curva de la **Regresión Logística** se mantiene ligeramente por encima de la de Random Forest en la mayor parte del recorrido, especialmente en el rango donde el FPR es bajo y el TPR es más alto. Esto se refleja en su AUC ligeramente superior (0.751 vs 0.748).

El modelo de Regresión Logística exhibe el mejor *trade-off* entre TPR y FPR al tener el AUC más alto y ubicarse en general más cerca de la esquina superior izquierda de la curva ROC.