Informe de Calibración de Modelos para la Predicción de Default Crediticio

Pablo Antonio García Pastor

26 de mayo de 2025

Índice

1. Introducción

La predicción del incumplimiento en pagos de tarjetas de crédito es un problema crítico para las instituciones financieras. Este informe presenta un análisis de datos y el desarrollo de modelos de machine learning para abordar este desafío, con un enfoque en la interpretabilidad y la evaluación crítica del rendimiento.

2. Análisis Exploratorio de Datos (EDA)

El análisis inicial se centró en comprender las características principales del conjunto de datos y la variable objetivo.

2.1. Distribución de Variables Clave

Se analizaron las distribuciones de las variables `LIMIT_BAL` (Límite de Crédito), `AGE` (Edad) y `default` (Incumplimiento de pago).

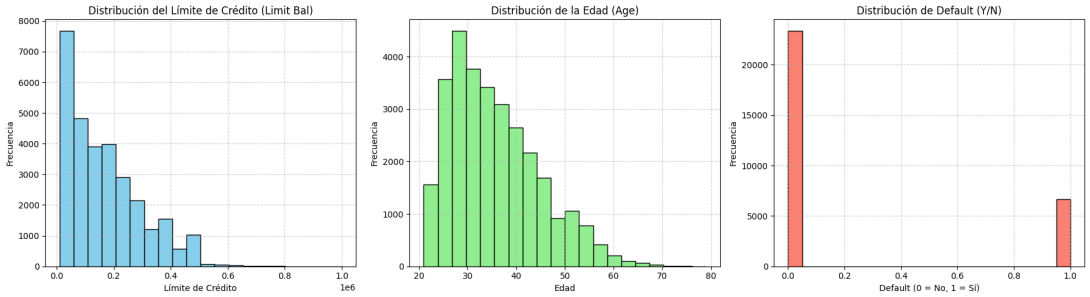

Figura 1: Distribución de Límite de Crédito, Edad y Default.

extttdefault): Existe un desbalance significativo. La clase "No Default"(0

2.1.1. Distribución del Límite de Crédito (Limit Bal)

- **Eje X (Límite de Crédito):** Muestra los montos del límite de crédito, variando desde 0 hasta 1,000,000 (1e6).
- **Eje Y (Frecuencia):** Indica el número de clientes que caen en cada rango de límite de crédito, alcanzando un máximo cercano a 8,000.
- **Forma y Observaciones:**
  - La distribución está **fuertemente sesgada a la derecha** (sesgo positivo).
  - La gran mayoría de los clientes posee límites de crédito bajos. La barra más alta, con una frecuencia superior a 7,500, se encuentra en el extremo inferior de los límites (probablemente entre 0 y 100,000).
  - A medida que el límite de crédito aumenta, la frecuencia de clientes disminuye drásticamente.
  - Existe una "cola larga" hacia la derecha, lo que significa que hay unos pocos clientes con límites de crédito muy altos, pero son una minoría.

2.1.2. Distribución de la Edad (Age)

- **Eje X (Edad):** Representa la edad de los clientes, en un rango aproximado de 20 a 80 años.
- **Eje Y (Frecuencia):** Muestra el número de clientes en cada rango de edad, con un pico superior a 4,000.
- **Forma y Observaciones:**
  - La distribución es **unimodal** y presenta un **ligero sesgo a la derecha**.
  - El grupo de edad más frecuente (la moda) se sitúa entre los 25 y 35 años, con la barra más alta alrededor de los 28-30 años (frecuencia aproximada de 4,300).
  - La frecuencia de clientes disminuye a medida que la edad aumenta, aunque de forma más gradual que en la distribución del límite de crédito.
  - Hay una concentración significativa de clientes en sus 20s, 30s y 40s.

2.1.3. Distribución de Default (Y/N)

- **Eje X (Default):** Variable categórica binaria, donde `0` representa "No Default"(el cliente no incumplió el pago) y `1` representa "Default"(el cliente sí incumplió).
- **Eje Y (Frecuencia):** Número de clientes en cada categoría de default. La barra para "No Default"supera los 20,000.
- **Forma y Observaciones:**
  - Es una distribución categórica con dos barras.
  - La barra correspondiente a `0` (No Default) es considerablemente más alta que la barra para `1` (Default).
  - Frecuencia aproximada de "No Default"(0): Entre 23,000 y 23,500 clientes.
  - Frecuencia aproximada de "Default"(1): Entre 6,500 y 7,000 clientes.

2.2. Relación entre Límite de Crédito y Default

Se utilizó un boxplot para visualizar la relación entre el límite de crédito y la ocurrencia de default.

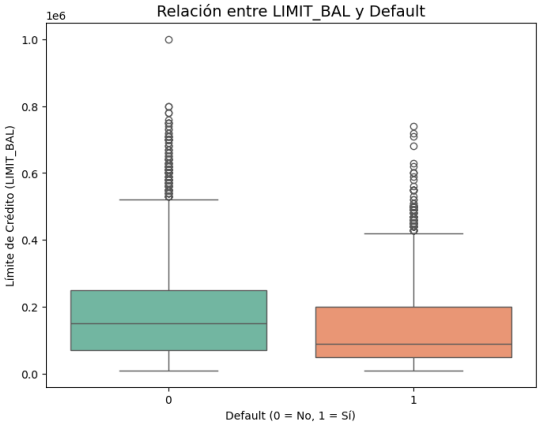

Figura 2: Relación entre Límite de Crédito (`LIMIT_BAL`) y Default.

El boxplot compara la distribución del **Límite de Crédito (`LIMIT_BAL`)** entre dos grupos de clientes:

- **Sin Default (0):** Muestran límites de crédito **significativamente más altos**. La mediana se sitúa alrededor de 150,000 y toda la distribución (incluyendo cuartiles y bigotes) se desplaza hacia valores mayores en comparación con el otro grupo.
- **Con Default (1):** Presentan límites de crédito **notablemente más bajos**. La mediana es de aproximadamente 80,000-90,000, y su distribución general de límites es inferior.

Conclusión Principal: Existe una clara tendencia: los **clientes que no incumplen pagos tienden a tener límites de crédito más elevados**, mientras que aquellos que **sí incumplen suelen tener límites más bajos**. Aunque hay solapamiento y valores atípicos en ambos grupos, la diferencia en las medianas y la posición general de las cajas indica que `LIMIT_BAL` es un factor distintivo entre los dos grupos.

3. Entrenamiento y Evaluación de Modelos

Se entrenaron y evaluaron tres modelos de clasificación para predecir el default.

3.1. Métricas de Evaluación

El rendimiento de los modelos se evaluó utilizando el reporte de clasificación (precisión, recall, F1-score) y el área bajo la curva ROC (AUC). La selección del mejor modelo se basó en un balance entre el AUC y el F1-score de la clase minoritaria (`Default = 1`).

3.2. Comparación de Modelos

La Tabla 1 resume los resultados obtenidos para cada modelo en el conjunto de prueba.

Cuadro 1: Tabla de Comparación de Modelos de Clasificación.

Modelo	AUC	F1 (Default)	Recall (Default)	Precisión (Default)	Accuracy	Balance (AUC*F1)
Random Forest	0.7544	0.4616	0.36	0.64	0.82	0.3482
Gradient Boosting	**0.7822**	**0.4625**	**0.35**	**0.67**	**0.82**	**0.3617**
Logistic Regression	0.7271	0.3515	0.24	0.69	0.81	0.2555

3.3. Interpretación Crítica de Resultados y Selección del Mejor Modelo

3.3.1. Análisis General del Rendimiento

- La **Accuracy** (Exactitud General) es alta para todos los modelos (0.81-0.82). Sin embargo, debido al conocido desbalance de clases en el dataset, esta métrica no es la principal para la decisión.
- El **AUC (Área Bajo la Curva ROC)**, que mide la capacidad discriminativa, es más alto para Gradient Boosting (0.7822), seguido por Random Forest (0.7544) y Logistic Regression (0.7271).

- El **F1-score para la clase `'Default'` (Clase 1)**, crucial para evaluar el rendimiento en la clase minoritaria, es más alto para Gradient Boosting (0.4625) y Random Forest (0.4616). Logistic Regression obtiene un F1-score menor (0.3515), indicando una mayor dificultad de este modelo para predecir correctamente los casos de default.

3.3.2. Desglose por Modelo

- **Random Forest Classifier:**
  - AUC: 0.7544
  - F1-score (Default): 0.4616
  - Recall (Default): 0.36 (identifica el 36 % de los defaults reales)
  - Precisión (Default): 0.64
- **Gradient Boosting Classifier:**
  - **AUC: 0.7822 (el más alto)**
  - **F1-score (Default): 0.4625 (el más alto)**
  - Recall (Default): 0.35
  - Precisión (Default): 0.67
- **Logistic Regression:**
  - AUC: 0.7271
  - F1-score (Default): 0.3515
  - Recall (Default): 0.24
  - Precisión (Default): 0.69 (aunque la precisión es alta, se logra a expensas de un recall muy bajo para la clase `'Default'`)

3.3.3. Selección del Mejor Modelo

El criterio de selección es el "balance entre AUC y F1-score de la clase minoritaria (`default = 1`)". Calculando el producto `AUC * F1-score (Clase 1)`:

- Random Forest: $0{,}7544 \times 0{,}4616 = 0{,}3482$
- **Gradient Boosting:** $0{,}7822 \times 0{,}4625 = 0{,}3617$
- Logistic Regression: $0{,}7271 \times 0{,}3515 = 0{,}2555$

El mejor modelo, según este criterio, es Gradient Boosting.
¿Por qué Gradient Boosting es el mejor modelo?

1. **Mayor AUC:** Con 0.7822, Gradient Boosting demuestra la mejor capacidad general para distinguir entre las clases de default y no default.
2. **Mejor "Balance" (AUC * F1-score Clase 1):** Alcanza el valor más alto (0.3617) en esta métrica combinada.

3. **Mejor F1-score para la Clase Minoritaria:** Gradient Boosting también presenta el F1-score más alto (0.4625) para la clase `'Default'`, indicando un mejor equilibrio entre precisión y recall para esta clase crucial.

4. **Buena Precisión para la Clase Minoritaria:** Su precisión para la clase `'Default'` (0.67) es la más alta entre los modelos con F1-scores competitivos.

En conclusión, Gradient Boosting se destaca como el modelo más equilibrado y con mejor rendimiento general para la tarea de predicción de default, según las métricas y el criterio de selección establecidos. No obstante, el recall para la clase `'Default'` (35 %) indica que todavía hay un margen importante para mejorar la identificación de los clientes que efectivamente incurrirán en impago.

4. Explicabilidad del Modelo (SHAP)

Se utilizó la librería SHAP (SHapley Additive exPlanations) para interpretar las predicciones del modelo Gradient Boosting.

4.1. Importancia Global de Características

El `summary plot` de SHAP (Figura 3) muestra la importancia global de cada característica y cómo sus valores impactan la predicción del modelo.

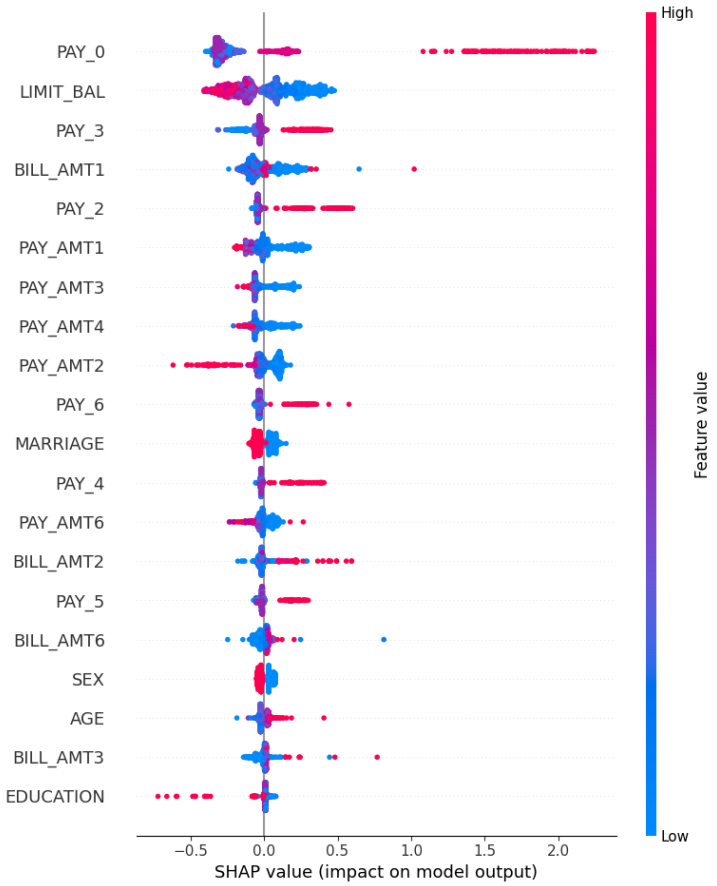

Figura 3: SHAP Summary Plot para el modelo Gradient Boosting.

Cómo leer el gráfico:

1. **Eje Vertical:** Las características están ordenadas de mayor a menor importancia (la de arriba es la más influyente). Los nombres que se muestran son los asignados a las características (ej. `PAY_0`, `LIMIT_BAL`).

2. **Eje Horizontal (SHAP value):** Representa el valor SHAP. Este valor indica la contribución de una característica a la predicción del modelo para una instancia específica (impacto en la salida del modelo).

   - **Valores Positivos (a la derecha de la línea central):** Empujan la predicción hacia la clase positiva (impago = 1). Un valor SHAP alto positivo significa que esa característica, con ese valor particular, aumenta fuertemente la probabilidad predicha de impago.
   - **Valores Negativos (a la izquierda de la línea central):** Empujan la predicción hacia la clase negativa (no impago = 0), disminuyendo la probabilidad de impago.

3. **Puntos:** Cada punto en una fila de característica representa una observación (un cliente) del subconjunto de datos de prueba utilizado. La posición horizontal del punto es su valor SHAP para esa característica.

4. **Color (Feature value):** El color de cada punto indica el valor original de la característica para esa observación, según la barra de color a la derecha:

   - **Rojo/Rosa (High):** Valores altos de la característica.
   - **Azul (Low):** Valores bajos de la característica.

Análisis del Gráfico (Figura 3):

- **Variables más importantes:**
  - Las características en la parte superior del gráfico son las más influyentes. En este caso, `PAY_0` es claramente la característica más importante para el modelo.
  - Le siguen en importancia `LIMIT_BAL`, `PAY_3`, `BILL_AMT1`, y `PAY_2`.
- **Impacto de los valores de las características en el riesgo de impago:**
  - `PAY_0`**:** Los valores altos de `PAY_0` (puntos rojos, indicando mayores retrasos en el pago más reciente) tienen valores SHAP consistentemente positivos y altos. Esto significa que un mayor retraso en el pago de septiembre aumenta significativamente el riesgo predicho de impago. Por el contrario, valores bajos de `PAY_0` (puntos azules, pago puntual o adelantado) tienden a tener valores SHAP negativos, disminuyendo el riesgo.
  - `LIMIT_BAL`**:** Para `LIMIT_BAL`, se observa una tendencia clara: valores bajos (puntos azules) tienen valores SHAP positivos, indicando que un límite de crédito bajo aumenta el riesgo de impago. Por otro lado, valores altos de límite de crédito (puntos rojos) tienen valores SHAP negativos, disminuyendo el riesgo.

- `PAY_3`, `PAY_2`, `PAY_6`, `PAY_4`, `PAY_5`: Similar a `PAY_0`, valores altos en estas características (retrasos en pagos de meses anteriores) tienden a empujar la predicción hacia el impago (valores SHAP positivos, puntos rojos a la derecha). Los pagos puntuales o adelantados (valores bajos, puntos azules) tienden a reducir el riesgo.
- **`BILL_AMT1` (y otras `BILL_AMTx`):** Para `BILL_AMT1`, valores altos (puntos rojos) se asocian predominantemente con valores SHAP positivos, sugiriendo que un alto monto de factura en septiembre aumenta el riesgo. Sin embargo, también hay algunos puntos rojos con SHAP negativo, lo que indica interacciones más complejas. Los valores bajos (azules) tienden a tener SHAP negativo o cercano a cero.
- **`PAY_AMT1` (y otras `PAY_AMTx`):** En general, para `PAY_AMT1`, valores bajos de pago (puntos azules) tienden a estar asociados con valores SHAP positivos (aumentan el riesgo de impago), mientras que valores altos de pago (puntos rojos) tienden a tener valores SHAP negativos (disminuyen el riesgo). Esto es intuitivo: pagar más reduce el riesgo.
- **`MARRIAGE`, `SEX`, `AGE`, `EDUCATION`:** Estas características demográficas tienen un impacto global menor en comparación con las variables de comportamiento de pago y límite de crédito. Para `EDUCATION`, por ejemplo, valores bajos (azul, posiblemente representando niveles educativos más bajos según la codificación original antes de la agrupación) tienen una ligera tendencia a SHAP negativo, mientras que algunos valores altos (rojo, posiblemente representando niveles educativos más altos o .ºtros") tienen SHAP positivo. Se necesitaría conocer la codificación exacta para una interpretación más precisa.

4.2. Explicación de una Predicción Individual

El `waterfall plot` de SHAP (Figura 4) desglosa la predicción para una observación específica, mostrando la contribución de cada característica.

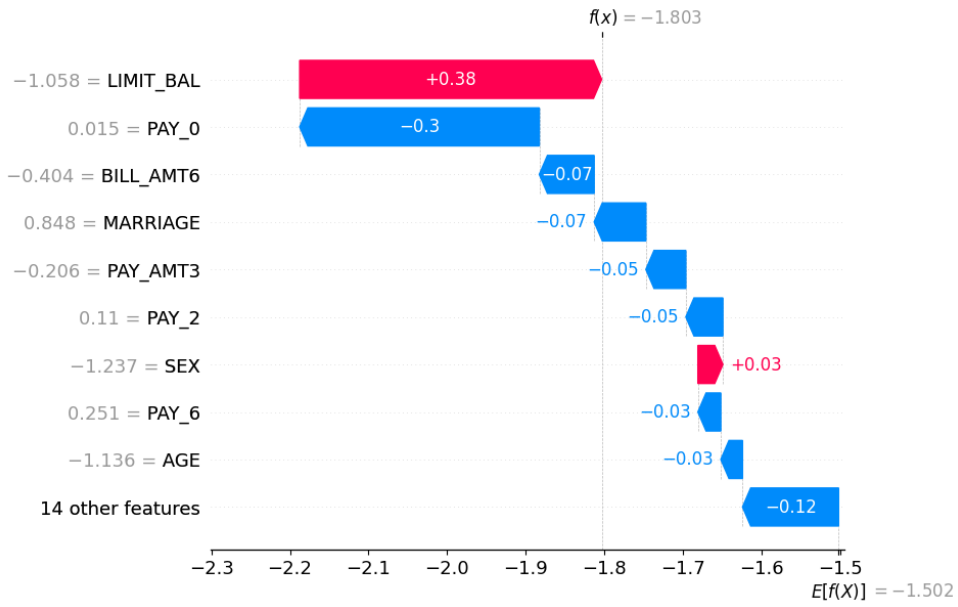

Figura 4: SHAP Waterfall Plot para una instancia de prueba (índice 0).

Cómo leer el gráfico:

1. **Eje Horizontal (salida del modelo):** Representa la salida del modelo (generalmente en escala de log-odds para modelos como Gradient Boosting o Logistic Regression).

2. **Eje Vertical (características):** Muestra las características, ordenadas por la magnitud de su contribución (valor SHAP absoluto) para *esta instancia*. A la izquierda de cada nombre de característica, se muestra el valor original (escalado) de esa característica para esta instancia.

3. **Valor Base `E[f(X)]`:** Es el valor SHAP esperado o la predicción promedio del modelo sobre el conjunto de datos de entrenamiento. Es el punto de partida de la explicación. En el gráfico de ejemplo, es `-1.502`.

4. **Barras (Flechas):** Cada barra representa la contribución (valor SHAP) de una característica *para este cliente específico*.

   - **Rojo (flecha hacia la derecha):** Indica que el valor de esa característica para este cliente empuja la predicción hacia arriba (aumenta la probabilidad de impago, o aumenta el log-odds).
   - **Azul (flecha hacia la izquierda):** Indica que el valor de esa característica para este cliente empuja la predicción hacia abajo (disminuye la probabilidad de impago, o disminuye el log-odds).
   - La longitud de la barra indica la magnitud de la contribución. El valor numérico sobre la barra es el valor SHAP de esa característica para esta instancia.

5. **Valor Final `f(x)`:** Es la predicción final del modelo (en log-odds) para esta instancia, resultado de sumar todas las contribuciones al valor base. En el gráfico de ejemplo, es `-1.803`. Un `f(x)` positivo generalmente corresponde a una predicción de clase 1 (Impago), y uno negativo a clase 0 (No Impago), donde el umbral es 0 en la escala log-odds.

Análisis del Gráfico (Figura 4 - Ejemplo): El gráfico muestra la explicación para una instancia específica.

- El modelo parte de una predicción base `E[f(X)] = -1.502`.
- **Características que Aumentan el Riesgo (Rojo):**
  - `LIMIT_BAL` = **-1.058:** Este valor (que es bajo, ya que los datos están escalados) tiene la mayor contribución positiva (+0.38), empujando significativamente la predicción hacia un mayor riesgo de impago. Esto es consistente con la observación global de que límites de crédito más bajos aumentan el riesgo.
  - `SEX` = **-1.237:** Este valor (posiblemente representando un género específico después de la codificación y escalado) tiene una contribución positiva (+0.03), aumentando ligeramente el riesgo.
- **Características que Disminuyen el Riesgo (Azul):**

- `PAY_0` = **0.015:** Este valor (cercano al promedio, posiblemente indicando pago puntual o un retraso muy pequeño) tiene la mayor contribución negativa (-0.30), disminuyendo considerablemente el riesgo de impago.
- `BILL_AMT6` = **-0.404** y `MARRIAGE` = **0.848** también contribuyen negativamente (-0.07 cada una), reduciendo el riesgo.
- Otras características como `PAY_AMT3`, `PAY_2`, `PAY_6`, y `AGE` tienen contribuciones negativas menores.
- Las "14 other featuresçombinadas también contribuyen negativamente (-0.12).

- **Predicción Final:** Sumando todas estas contribuciones al valor base, la predicción final para esta instancia es `f(x) = -1.803`. Como este valor es negativo (y más negativo que el valor base), el modelo predice "No Impago"(clase 0) para este cliente. La información del notebook indica que para la instancia 0, el valor real y el predicho fueron ambos 0 (No Default), lo cual es consistente con este `f(x)` negativo.

5. Reflexión y Evaluación Crítica

El modelo **Gradient Boosting** fue el de mejor rendimiento general. Sin embargo, una limitación significativa es su bajo recall (aproximadamente 34 %) para la clase `'Default'`, lo que significa que no identifica a la mayoría de los clientes que realmente impagarán. Esta debilidad se atribuye principalmente al desbalance de clases en el dataset.

Recomendaciones para Mejoras Futuras:

- **Aplicar Técnicas de Balanceo de Clases:** Considerar el uso de SMOTE (Synthetic Minority Over-sampling Technique) en el conjunto de entrenamiento para generar ejemplos sintéticos de la clase minoritaria y ayudar al modelo a aprender mejor sus patrones.
- **Ajustar el Umbral de Decisión:**** Explorar la disminución del umbral de clasificación (actualmente 0.5) para incrementar el recall de la clase `'Default'`, aceptando un posible aumento en los falsos positivos. La elección del umbral óptimo dependerá de un análisis coste-beneficio.
- **Penalización de Clases/Pesos:**** Asignar un mayor peso a los errores en la clase minoritaria durante el entrenamiento del modelo (e.g., usando `scale_pos_weight` en Gradient Boosting).

Estas estrategias podrían mejorar la capacidad del modelo para identificar clientes con riesgo real de impago, haciéndolo más útil para la toma de decisiones de negocio.