Informe Explicativo: Modelo de Clasificación (Parcial 2 - Finanzas Descentralizadas)

1. Contexto del proyecto

Este trabajo corresponde al **segundo punto del Parcial de Finanzas Descentralizadas IV**, cuyo objetivo fue construir, evaluar y justificar un **modelo de machine learning predictivo** aplicado al contexto crediticio. El modelo busca anticipar el riesgo de mora o impago a partir de información histórica de clientes. Se evalúa no solo por su exactitud, sino también por: - **Recall**: capacidad para identificar correctamente a los clientes con mora (minimizar falsos negativos). - **Sobreajuste**: diferencia entre desempeño en entrenamiento/validación y en test. - **Valor agregado de los intereses a un año**: impacto financiero de aplicar correctamente el modelo.

2. Preparación de datos y preprocesamiento

- Se partió de una base de 40.000 observaciones con una variable objetivo binaria (`TARGET`).
- Se realizó una división estratificada en proporciones **70% entrenamiento / 15% validación / 15% prueba**.
- Se identificaron variables numéricas y categóricas:
- Numéricas: imputadas con la **mediana**.
- Categóricas: imputadas con la **moda** y transformadas mediante **OneHotEncoder**.
- Se calculó el `scale_pos_weight` ≈ 11.7 para compensar el fuerte desbalance entre clases (pocos casos con mora).

3. Modelos implementados

Se implementaron tres enfoques comparativos: 1. **Regresión Logística (baseline lineal)** – simple, interpretable, con `class_weight='balanced'`. 2. **XGBoost** – modelo basado en árboles con ponderación de clases, optimizado para `AUC-PR`. 3. **LightGBM** – alternativa más ligera, también optimizada con el mismo objetivo.

> En este notebook se presentan los resultados del modelo **Logistic Regression**, elegido como referencia principal para evaluar el rendimiento mínimo esperado.

4. Selección de umbral (threshold)

Dado el desbalance, se eligió el umbral **0.299**, ajustado para **maximizar el Recall** manteniendo una **Precisión mínima de 10%**. Esto permite priorizar la detección de morosos, incluso si implica más falsos positivos (criterio conservador usado en créditos).

5. Resultados del modelo (Logistic Regression)

Métrica	VALIDACIÓN	TEST
AUC-PR	0.1870	0.2079

Métrica	VALIDACIÓN	TEST
AUC-ROC	0.7171	0.7399
Recall	0.8962	0.9195
Precision	0.0999	0.1037
F1 Score	0.1799	0.1863
Threshold	0.299	0.299

Matriz de confusión (Test)

```
[[1776, 3752],
 [  38,  434]]
```

Interpretación: - El modelo **detecta más del 91%** de los clientes con riesgo de mora (alta sensibilidad), cumpliendo el criterio principal del parcial. - Sin embargo, la **precisión baja (~10%)** indica que de cada 10 alertas, solo 1 resulta ser mora real (esperable en datasets desbalanceados). - El **AUC-PR** de 0.21 es competitivo frente a la prevalencia (~8%), lo que sugiere valor predictivo real.

6. Análisis de sobreajuste

El desempeño en validación y prueba es **muy similar**, lo que indica **ausencia significativa de sobreajuste**. - AUC-PR y AUC-ROC mejoran ligeramente en test, lo que puede deberse al muestreo estratificado. - No hay evidencia de que el modelo haya memorizado el entrenamiento.

7. Valor agregado de los intereses a un año

Para estimar el impacto financiero: - El modelo predice correctamente 434 morosos (TP) y evita prestar a clientes que habrían generado pérdidas. - Suponiendo un **interés anual del 20%** y una pérdida promedio de **X unidades por mora**, la reducción potencial de pérdidas sería:

$$Valor agregado \approx 434 * (pérdida evitada por mora) - 3752 * (interés no ganado por falsos positivos)$$

- La política de negocio puede ajustar el umbral para optimizar este balance (por ejemplo, buscando un Recall del 85% con mayor precisión para maximizar beneficios netos).

8. Conclusiones generales

- El modelo baseline logra **excelente Recall (0.92)** y **buen AUC-PR (0.21)**, demostrando capacidad real de identificación temprana de riesgo crediticio.
- No se observan signos de sobreajuste.
- Los falsos positivos pueden ser mitigados ajustando el umbral o combinando este modelo con reglas de negocio.

- El enfoque puede ampliarse con modelos más complejos (XGBoost o LightGBM) para mejorar la precisión sin sacrificar Recall.

9. Próximos pasos recomendados

1. Ajustar modelos no lineales (XGBoost/LightGBM) con early stopping.
2. Calibrar probabilidades con `CalibratedClassifierCV` para mejorar la interpretación de riesgo.
3. Simular el impacto económico real de distintas políticas de umbral (curvas costo-beneficio).
4. Documentar la trazabilidad del modelo para presentación en el informe final.

Resumen ejecutivo:

> Se desarrolló un pipeline completo de clasificación binaria orientado a riesgo de mora. El modelo lineal baseline obtuvo un Recall de 0.92 sin sobreajuste y un AUC-PR de 0.21. Esto permite detectar preventivamente casi todos los casos de mora con una precisión modesta, cumpliendo los criterios principales de evaluación del parcial y sentando las bases para optimización futura mediante modelos de árboles.