Memoria de la Práctica Final: Explicabilidad y Ética en IA

Jorge Gómez Azor

Diciembre 2025

1. Introducción

1.1. Descripción del problema

Esta práctica intenta servir de ayuda para entender las decisiones de un modelo de Credit Scoring, cuyo objetivo es predecir la probabilidad de que un solicitante de préstamo cumpla con sus obligaciones de pago. Se trata de un problema de clasificación binaria donde categorizamos a los clientes como "Good"(cumplidores) o "Bad"(morosos). En el sector financiero, esta tarea es crítica para minimizar el riesgo de impago y es algo que se ve en el día a día.

1.2. Dataset

Se utiliza el dataset **HELOC (Home Equity Line of Credit)**, proporcionado por FICO para su Explainable Machine Learning Challenge. El dataset contiene información real de solicitudes de líneas de crédito.

- **Tamaño:** Aproximadamente 10000 registros tras la limpieza inicial.
- **Tipos de datos:** Variables numéricas enteras y una variable objetivo categórica.

1.3. Stakeholders

El principal stakeholder es el departamento de riesgos de una entidad bancaria, así como el propio cliente solicitante. La explicabilidad es fundamental en este escenario por dos razones en mi opinión:

- **Regulación (GDPR):** Esta regulación obliga a los bancos a justificar por qué se deniega un crédito.
- **Accionabilidad:** Permite dar recomendaciones al cliente sobre qué factores mejorar para obtener crédito en el futuro.

1.4. Diccionario de variables

A continuación, se describen las variables del dataset utilizadas para el entrenamiento del modelo:

Variable	Descripción
RiskPerformance	Variable objetivo (Good/Bad).
ExternalRiskEstimate	Puntuación consolidada de agencias externas de crédito.
MSinceOldestTradeOpen	Meses transcurridos desde la primera operación comercial.
MSinceMostRecentTradeOpen	Meses desde la operación comercial más reciente.
AverageMInFile	Antigüedad media de las cuentas en el historial.
NumSatisfactoryTrades	Número de operaciones comerciales satisfactorias.
NumTrades60Ever2DerogPubRec	Operaciones con más de 60 días de morosidad.
NumTrades90Ever2DerogPubRec	Operaciones con más de 90 días de morosidad.
PercentTradesNeverDelq	Porcentaje de operaciones que nunca han tenido mora.
MSinceMostRecentDelq	Meses desde la última morosidad registrada.
MaxDelq2PublicRecLast12M	Máxima morosidad en registros públicos en los últimos 12 meses.
MaxDelqEver	Máxima morosidad histórica del cliente.
NumTotalTrades	Número total de operaciones comerciales.
NumTradesOpeninLast12M	Número de operaciones abiertas en el último año.
PercentInstallTrades	Porcentaje de préstamos a plazos.
MSinceMostRecentInqexcl7days	Meses desde la última consulta de crédito (excluyendo última semana).
NumInqLast6M	Número de consultas de crédito en los últimos 6 meses.
NetFractionRevolvingBurden	Ratio de utilización de líneas de crédito revolving.
NetFractionInstallBurden	Ratio de utilización de préstamos a plazos.
NumRevolvingTradesWBalance	Número de cuentas revolving con saldo pendiente.
NumInstallTradesWBalance	Número de préstamos a plazos con saldo pendiente.
NumBank2NatlTradesWHighUtil	Operaciones bancarias nacionales con alta utilización.
PercentTradesWBalance	Porcentaje de todas las cuentas con saldo pendiente.

Cuadro 1: Variables del dataset HELOC.

2. Preprocesamiento y Entrenamiento

2.1. Limpieza de datos y gestión de valores especiales

El dataset presenta una particularidad en sus valores numéricos: la presencia de códigos de error específicos de FICO (-7, -8, -9). Estos valores no representan magnitudes, sino situaciones como çuenta no utilizada.º ïnformación no disponible".

En el preprocesamiento, se han tratado estos valores para evitar que el modelo los interprete como medidas reales. Sin embargo, vi que los -7 y -8 son errores puntuales y no afectaban mucho al modelo y, si me quitaba todos el dataset se quedaba muy pequeño por lo que opté por eliminarme solo las filas con el código de error -9 (que son errores de todos los valores a la vez).

2.2. Análisis del balanceo de clases

Tras la limpieza inicial, la distribución de clases resultó ser:

- **Bad (1):** 52.19 %
- **Good (0):** 47.81 %

Al estar el dataset prácticamente balanceado, no ha sido necesario aplicar técnicas de remuestreo.

2.3. Configuración del Modelo y Resultados

Se ha seleccionado un clasificador XGBoost por su excelente rendimiento con datos tabulares y su capacidad para manejar relaciones no lineales entre variables que en este caso eran finan-

cieras. El modelo fue evaluado con el split de test (20 % de los datos), obteniendo los siguientes resultados:

Métrica	Valor
AUC Score	0.7908
Accuracy	0.7161

Cuadro 2: Métricas de rendimiento del modelo final.

El AUC de 0.79 indica una capacidad sólida de discriminación entre clientes solventes y de riesgo. Según el classification report, el modelo muestra un equilibrio notable entre precisión y recall, siendo ligeramente más robusto en la detección de casos negativos (Recall de 0.76 para la clase "Bad"), lo cual es positivo en un entorno bancario para minimizar el riesgo de impago.

3. Interpretabilidad Global (SHAP)

Para entender mejor el comportamiento general del modelo, se ha utilizado el *Summary Plot* de SHAP. Este gráfico permite visualizar no solo la importancia de las variables, sino también la dirección del impacto de sus valores en la predicción final.

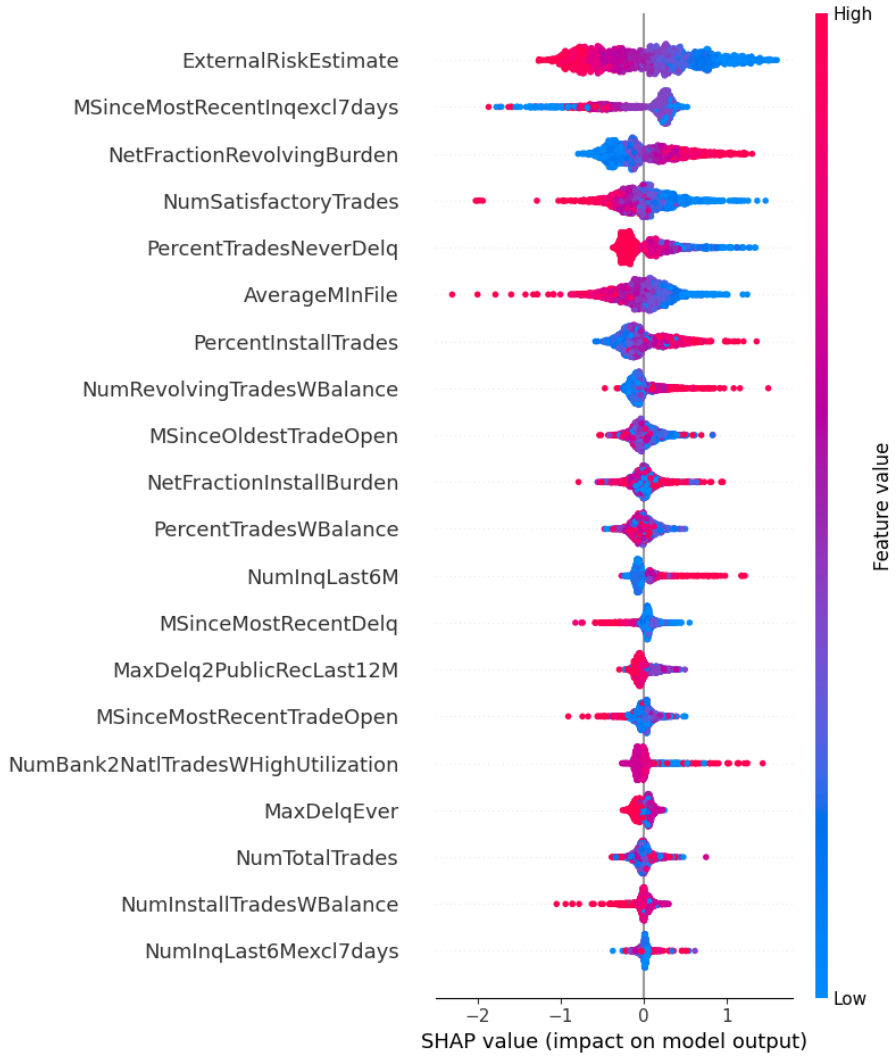

Figura 1: SHAP Summary Plot: Impacto de las características en el modelo.

A partir de estos resultados globales, quiero resaltar algunas conclusiones:

- **Dominancia del score externo:** La variable *ExternalRiskEstimate* es, con diferencia, la más influyente. Se observa que valores bajos (puntos azules) tienen un impacto positivo alto en el SHAP, lo que aumenta la probabilidad de ser clasificado como Bad. Por el contrario, valores altos (puntos rojos) reducen drásticamente el riesgo.
- **Comportamiento de las consultas:** La variable *MSinceMostRecentInqexcl7days* muestra que la falta de actividad reciente en consultas de crédito se asocia con un menor riesgo. En cambio, valores cercanos a cero (solicitudes muy recientes) aumentan algo la probabilidad de impago.
- **Uso del crédito revolving:** Un alto ratio de utilización en cuentas revolving (*NetFractionRevolvingBurden* en rojo) correlaciona directamente con el desplazamiento hacia la clase Bad, validando la lógica financiera de que un cliente excesivamente endeudado es más probable que no te pague.
- **Consistencia del modelo:** La distribución de los puntos en el gráfico confirma que el modelo ha aprendido patrones monótonos y lógicos, lo que aumenta la confianza de los stakeholders en sus decisiones.

4. Interpretabilidad Local (SHAP)

Se ha analizado la predicción individual de un perfil con riesgo extremo para entender los motivos de su clasificación. El modelo asigna a este cliente una log-probabilidad de 1.379, situándolo claramente en la clase Bad.

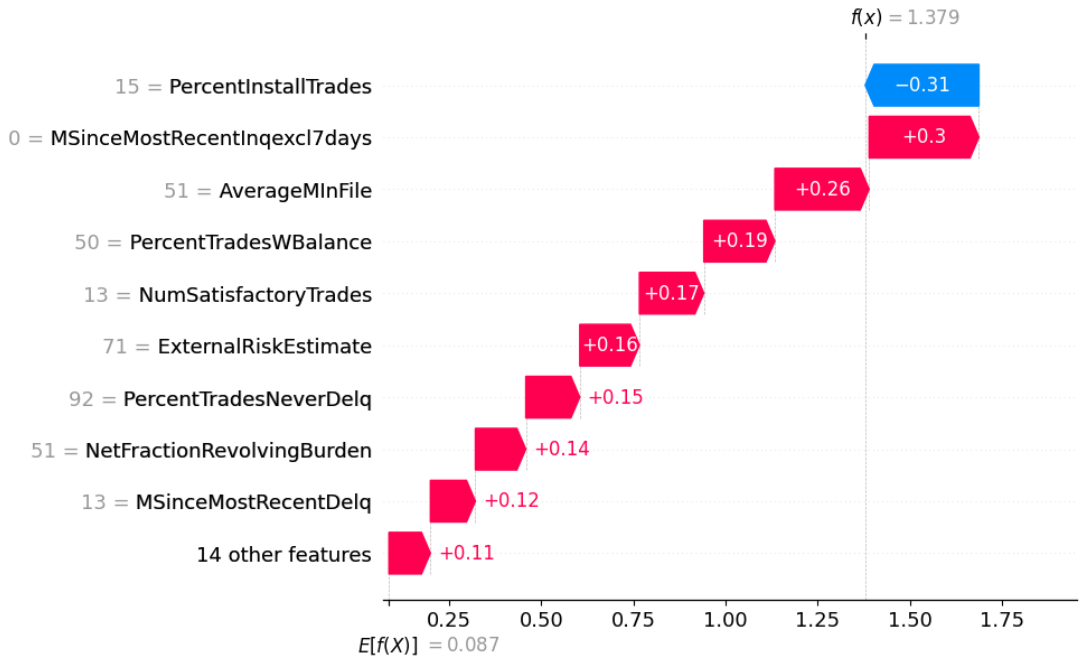

Figura 2: SHAP Waterfall Plot: Desglose de la predicción para un cliente de alto riesgo.

Las conclusiones principales de este análisis son:

- **Factores críticos:** La reciente actividad crediticia (MSinceMostRecentInq = 0) y un historial de cuentas joven (AverageMInFile = 51) son los mayores agravantes, sumando +0.30 y +0.26 al riesgo respectivamente.

- **Efecto acumulativo:** A diferencia de otros perfiles, casi todas las variables del cliente empujan la predicción hacia el impago.
- **Mitigación insuficiente:** La baja proporción de préstamos a plazos (PercentInstallTrades = 15) resta -0.31 al riesgo, pero este impacto positivo es irrelevante frente al peso del resto de indicadores negativos.

5. Análisis de Contrafactuales (DiCE)

Para complementar el análisis, se ha utilizado la librería **DiCE** (*Diverse Counterfactual Explanations*). Esta técnica no explica por qué el modelo tomó una decisión, sino qué cambios mínimos serían necesarios en los datos del cliente para que el modelo cambie su predicción de **Bad (1)** a **Good (0)**.

Cabe mencionar que no todas las variables pueden ser modificadas ya que eso sería irreal, y en el notebook menciono y explico brevemente porque elijo esas variables (NetFractionRevolvingBurden, NumRevolvingTradesWBalance, NumInqLast6Mexcl7days, NumSatisfactoryTrades, PercentTradesWBalance, NumInstallTradesWBalance, NumTradesOpeninLast12M y PercentInstallTrades).
Estos son algunos ejemplos para el mismo ejemplo que tomé en la interpretabilidad local:

#	NumSatisfactoryTrades	NumInstallTradesWBalance	PercentTradesWBalance	...	Target
Org.	13	1	50	...	Bad
1	79	1	50	...	Good
2	13	15	10	...	Good
3	74	1	50	...	Good
4	70	1	50	...	Good
5	40	1	5	...	Good

Cuadro 3: Ejemplos contrafactuales generados por DiCE para revertir la clasificación.

5.1. Análisis de las soluciones propuestas

Las soluciones sugeridas por DiCE muestran caminos claros para que el cliente mejore su perfil crediticio:

- **Aumento de operaciones satisfactorias:** La vía más recurrente implica incrementar drásticamente el número de cuentas cerradas con éxito (pasando de 13 a valores entre 40 y 79), lo que demuestra estabilidad financiera a largo plazo.
- **Reducción del porcentaje de saldo pendiente:** El modelo sugiere que bajar el porcentaje de cuentas con balance actual del 50 % a valores mínimos (5 % o 10 %) es suficiente para reclasificar al cliente como "Good", incluso sin alterar radicalmente otras variables.
- **Gestión de préstamos activos:** Se observa que el modelo acepta perfiles con más préstamos a plazos siempre que el uso del crédito total sea bajo.

6. Sanity Checks

Para garantizar que las explicaciones proporcionadas por SHAP son de verdad útiles y coherentes se han realizado dos pruebas de validación.

6.1. Randomización de parámetros (Model Parameter Randomization Test)

Este test consiste en comparar las explicaciones del modelo entrenado frente a un modelo con pesos aleatorios. Si las explicaciones fueran independientes de los parámetros aprendidos, la técnica de XAI perdería fiabilidad.

- **Resultado:** Se obtuvo una correlación de -0.1913.

Al ser una correlación cercana a cero y negativa, el test se considera superado. Esto confirma que las explicaciones de SHAP cambian drásticamente cuando el modelo no ha aprendido, demostrando que en el modelo real las explicaciones sí dependen de los pesos entrenados.

6.2. Importancia por permutación (Feature Perturbation)

Se evaluó la sensibilidad del modelo ante la variable más importante (*ExternalRiskEstimate*) permutando sus valores aleatoriamente para romper su relación con el objetivo.

- **Baseline AUC:** 0.7964.
- **AUC tras permutar:** 0.7663.
- **Caída de performance:** 0.0301.

La degradación del 3 % en el rendimiento confirma que el modelo depende genuinamente de esta variable para discriminar el riesgo. Sin embargo, me sorprende que la caída sea tan baja siendo esta la variable más importante. Aunque al haber tantas variables no me parece tampoco incoherente.

7. Conclusiones

7.1. Acciones e insights: Mejora del modelo y negocio

A la vista de los resultados, el modelo tiene un margen de mejora claro si logramos tratar mejor el ruido de las variables con muchos valores que no aportan información (códigos -7, -8), que actualmente limitan el AUC a 0.79. Como recomendaciones al negocio siento que el análisis de contrafactuales abre un nuevo mundo dentro de la IA. Permite a los bancos usar modelos más complejos con explicaciones al cliente de igual o mejor calidad.

7.2. Limitaciones, riesgos y trabajo futuro

El mayor riesgo que he detectado es la dependencia del modelo hacia factores históricos rígidos. Al hacer los Sanity Checks, la correlación de -0.19 confirma que el modelo es robusto, pero también que es muy sensible a cualquier cambio en los parámetros, lo que podría hacerlo inestable ante crisis económicas no reflejadas en el histórico de FICO. Una limitación ética importante es que el modelo penaliza el comportamiento reciente (consultas de crédito) de forma muy agresiva, lo que podría generar un sesgo contra personas que simplemente están comparando opciones bancarias. Como trabajo futuro, sería clave implementar un sistema que ayude a corregir estos sesgos, limitando la capacidad de decisión del modelo a variables más objetivas. También sería interesante ver variantes a DiCE como otras opciones para ofrecer soluciones al cliente.