Universidad ICESI

45 años

Modelo predictivo para el análisis del riesgo de crédito

Grupo 7 - Proyecto de Innovación I

David Crespo - Julian Espinosa - Jhon Cuervo

Maestría en Inteligencia Artificial Aplicada

28-11-2025

Se evaluará utilizando la rúbrica completa:

- **Introducción y planteamiento del problema**
- **Metodología**: claridad y coherencia en la metodología aplicada
- **Resultados y discusión**: análisis de los resultados obtenidos y calidad de la discusión
- **Conclusiones**: aportes relevantes y alineación con los objetivos planteados
- **Referencias**: uso correcto de fuentes relevantes
- **Presentación en video**: claridad, síntesis y calidad visual del contenido

Contexto y motivación

Agenda

1. Planteamiento y Definición del Problema
2. Metodología
3. Discusiones y Resultados

1. Planteamiento del Problema

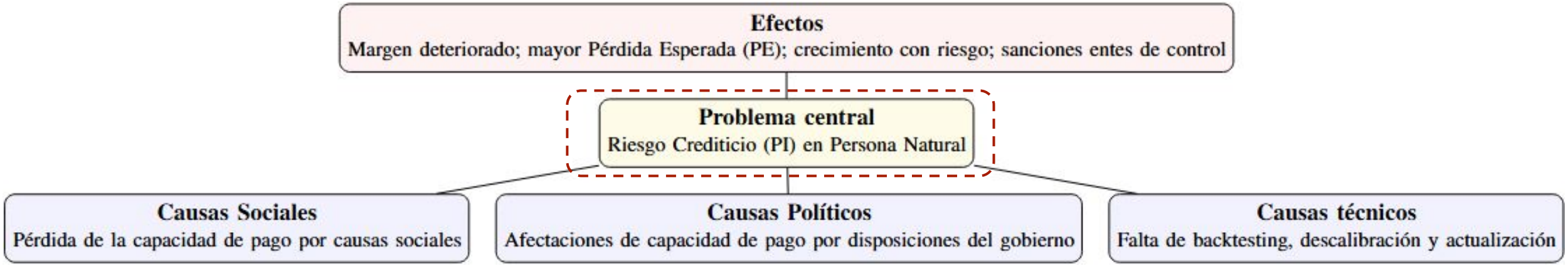
Planteamiento del Problema - Motivación

1. Crecimiento de la cartera de **consumo** y necesidad de gestionar adecuadamente el riesgo de crédito.

2. Regulación de la Superintendencia Financiera que exige modelos robustos de **probabilidad de incumplimiento**.

3. Oportunidad de **optimización** y **calibración** de los modelos incorporando técnicas de aprendizaje automático.

Planteamiento del Problema - Árbol del Problema



¿Qué es el riesgo de crédito?

Delimitación del problema - Definición de riesgo de crédito

"Es la posibilidad de que la entidad incurra en pérdidas y disminuya el valor de sus activos como consecuencia de que un deudor o contraparte incumpla sus obligaciones." [1]

Pérdida Esperada (PE) = Probabilidad de incumplimiento (PI) x Exposición del Activo (ExA) x Pérdida esperada de valor del activo dado el incumplimiento (PDI)

"Créditos de consumo que se encuentren en mora mayor a 90 días, o que siendo reestructurados incurran en mora mayor o igual a 60 días." - Buenos y Malos clientes

1 Tomado de: CAPÍTULO XXXI SISTEMA INTEGRAL DE ADMINISTRACIÓN DE RIESGOS (SIAR) - Superintendencia Financiera de Colombia

2. Metodología - EDA

Descripción de los datos

1 La base de datos utilizada es la base de datos **German Credit**, que ha sido utilizada en diferentes proyectos para estimar el riesgo crediticio.

2 La base cuenta con mil **(1.000) observaciones**, **21 variables** diferentes, de las cuales 14 son categóricas y las demás son variables numéricas.

3 Es una base que no cuenta con valores nulos ni valores duplicados y tiene un **desbalance de 70%-30%** en la variable respuesta.

Descripción de los datos

RESUMEN DEL ANÁLISIS EXPLORATORIO DE LAS VARIABLES PRINCIPALES

Variable	Tipo	Comportamiento observado	Implicaciones para el modelo
credit_amount	Numérica	Sesgo marcado a la derecha; pocos créditos de monto alto.	Transformación (p. ej. log) o winsorización para reducir efecto de outliers.
duration	Numérica (ordinal)	Plazos discretos (12, 24, 36, 48, etc.).	Tratar como ordinal; posible agrupación supervisada según riesgo.
age	Numérica	Unimodal, centrada en 30–40 años, con cola hacia edades altas.	Revisar relación no lineal; considerar bins o términos no lineales.
installment_rate	Ordinal discreta	Concentración en pocas categorías (1–4).	Usar como ordinal; buscar tendencia monotónica con el riesgo.
residence_since	Ordinal discreta	Mayor permanencia en algunas categorías; indicador de estabilidad.	Mantener como ordinal; evaluar asociación con incumplimiento.
existing_credits	Discreta	Mayoría con 1–2 créditos vigentes.	Captura endeudamiento previo; usar como categórica discreta.
num_dependents	Discreta	Muy concentrada en un solo valor.	Baja variabilidad; posible aporte sólo en interacciones.

Variable	Descripción / Significado
checking_status	Estado de la cuenta corriente del cliente (nivel de liquidez).
duration	Duración del crédito solicitado (en meses).
credit_history	Historial crediticio del cliente: comportamiento de pago previo.
purpose	Propósito o destino del crédito (auto, muebles, educación, etc.).
credit_amount	Monto total del crédito solicitado (en marcos alemanes, DM).
savings	Nivel de ahorros del cliente (<100, 100-500, >=1000 DM, o sin cuenta).
employment	Antigüedad en el empleo actual (desempleado, <1, 1–4, 4–7, >=7 años).
installment_rate	Porcentaje del ingreso disponible destinado a la cuota mensual (1 a 4).
personal_status	Estado civil y género del cliente (ej.: hombre soltero, mujer casada).
other_debtors	Si existen otros deudores o garantes asociados al crédito.
residence_since	Años de residencia en el domicilio actual.
property	Tipo de propiedad o activo que posee (bienes raíces, seguro, auto, etc.).
age	Edad del cliente (en años).
other_installment_plans	Otros planes de pago o créditos existentes (banco, tiendas, ninguno).
housing	Situación habitacional (alquiler, propia, vive gratis).
existing_credits	Número de créditos o préstamos en curso.
job	Tipo de empleo o nivel ocupacional (calificado, autónomo, gerente, etc.).
num_dependents	Número de dependientes económicos (por lo general hijos).
telephone	Si posee teléfono registrado a su nombre.
foreign_worker	Si es trabajador extranjero (yes/no).
target	Variable objetivo: clasifica al cliente como 'good' (bajo riesgo) o 'bad' (alto riesgo).

2. Metodología - Selección de variables

1. Se seleccionan 10 variables con mayor correlación con correlaciones entre el 10% y el 32% con respecto a la variable objetivo.

2. Las variables seleccionadas son variables con información del nivel de liquidez del cliente, ahorros, historial de pagos pasados, entre los más relevantes.

Metodología - Modelos

Se probaron diferentes modelos para estimar el riesgo crediticio

Regresión logística

Para regresión logística se exploraron alternativas como Lasso, Ridge, Reg. Logística balanceada y optimizada y Reg. Logística con Smoote.

Árboles de decisión

En árboles de decisión capturando relaciones no lineales y reglas de decisión explícitas fáciles de interpretar.

XGBoost

Para XGBoost se exploraron alternativas como el balanceo de clases y optimizando el mejor umbral. También se exploró el modelo LightGBM que maneja datos desbalanceados

Gradient Boosting

Se explora el modelo gradient boosting con el objetivo de encontrar un mejor recall en comparación a los modelos ya evaluados.

Modelos implementados y desempeño

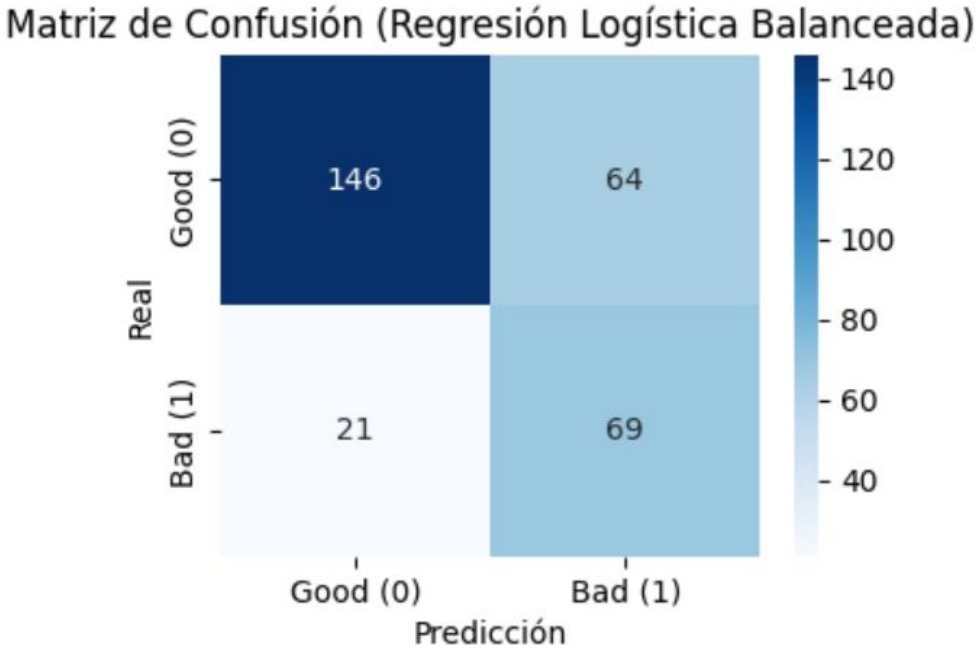
- Regresión logística como modelo de referencia por su interpretabilidad y amplio uso en riesgo de crédito.
- El modelo logístico sin balanceo logró un buen desempeño global (Accuracy ≈ 0.77), pero con capacidad limitada para detectar clientes de alto riesgo (Recall ≈ 0.51).
- El modelo de Árbol de decisión captura relaciones no lineales y ofrece reglas explícitas, pero obtuvo un rendimiento inferior (Accuracy ≈ 0.65, Recall ≈ 0.43).
- Los resultados resaltan la necesidad de balancear las clases y ajustar umbrales de decisión para mejorar la detección de clientes con alta probabilidad de incumplimiento.

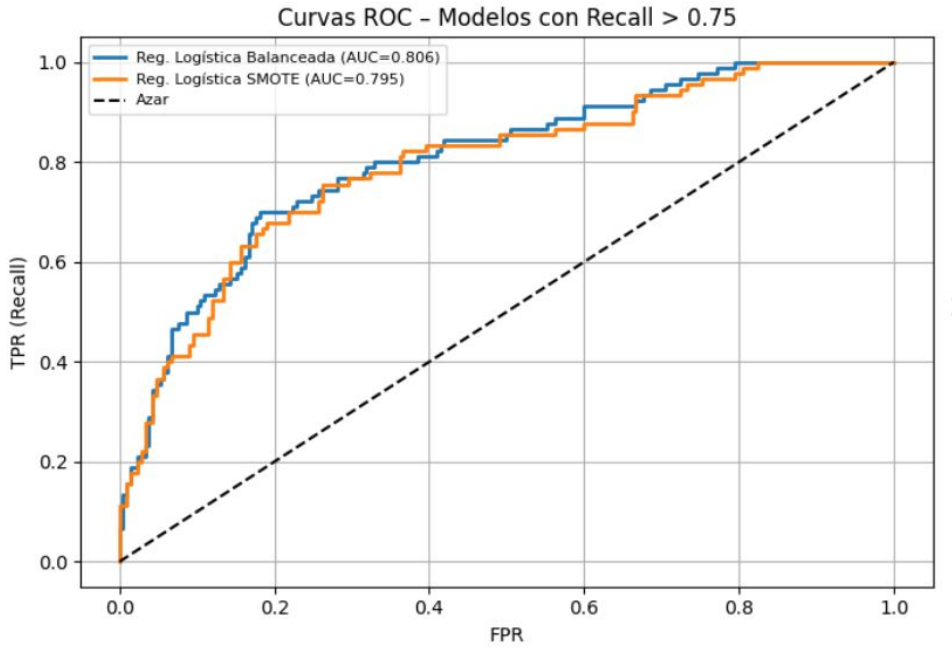
3. Resultados - Interpretación e Interpretaciones

Modelos implementados y desempeño

	Modelo	Accuracy	Precision	Recall	F1	AUC
0	Reg. Logística Balanceada	0.716700	0.518800	0.766700	0.618800	0.806200
1	Reg. Logística SMOTE	0.740000	0.548400	0.755600	0.635500	0.794700
2	Reg. Log. Bal. Opt. (u=0.59)	0.776700	0.611700	0.700000	0.652800	0.806200
3	Random Forest	0.730000	0.541300	0.655600	0.593000	0.774800
4	XGBoost Bal. Opt. (u=0.48)	0.743300	0.565700	0.622200	0.592600	0.752200

Modelos implementados y desempeño





Modelos implementados y desempeño

1. La probabilidad de default se reduce en un 48% si el cliente no tiene cuenta corriente.

2. La probabilidad de default aumenta en 37% por cada aumento en la duración del crédito.

	Variable	Coeficiente	Odds Ratio
2	duration	0.320157	1.377344
4	installment_rate	0.277757	1.320166
3	credit_amount	0.225854	1.253392
5	savings_< 100 DM	0.201852	1.223667
1	checking_status_< 0 DM	0.130855	1.139803
8	credit_history_no credits taken / all paid duly	0.122997	1.130881
6	savings_unknown / no savings account	-0.199410	0.819214
9	purpose_radio/television	-0.242580	0.784601
10	purpose_car (used)	-0.287924	0.749819
7	credit_history_critical account / other credit...	-0.389940	0.677098
0	checking_status_no checking account	-0.645884	0.524199

3. Conclusiones

1. Se evaluaron 14 modelos diferentes, encontrando al logístico balanceado como el modelo con mejores métricas, con un **recall del 76,7%.**

2. Uno de los puntos a destacar es la mejora de los resultados de los modelos al tratar el desbalanceo de clases, mejorando hasta en **20 puntos porcentuales el recall.**

3. Según el estado del arte, un modelo de predicción de riesgo crediticio es **considerado aceptable** con un **recall a partir del 60%.**

Universidad ICESI

45 años

Gracias!!