Proyecto aplicado: Sistema de pricing ajustado al riesgo para pymes que otorgan financiación

Introducción

Este documento propone una alternativa práctica y viable al enfoque tradicional de scoring bancario. En lugar de reproducir modelos complejos ya utilizados por grandes entidades financieras, se plantea construir un MVP que permita a pequeñas y medianas empresas calcular un tipo de interés personalizado en función del riesgo individual del cliente. Esta solución está diseñada para implementarse en solo 4 semanas y orientada a pymes como agencias inmobiliarias, concesionarios o tiendas que ofrecen financiación directa.

1. Contexto y enfoque del proyecto

El proyecto original planteaba construir un sistema de scoring de riesgo para entidades bancarias, utilizando modelos predictivos avanzados que estimaran la probabilidad de impago (PD), la exposición al impago (EAD) y la pérdida en caso de impago (LGD). Aunque ambicioso, este enfoque presenta dos desafíos importantes:

- Es complejo de implementar en solo 4 semanas.
- Reproduce sistemas ya establecidos en banca tradicional, sin un valor diferencial claro.

La propuesta alternativa es desarrollar una herramienta capaz de estimar la probabilidad de impago de un cliente utilizando datos básicos y, a partir de ella, recomendar un tipo de interés mínimo que cubra el riesgo. Este enfoque es original, aplicable en entornos reales de pymes, y puede desarrollarse de forma ágil.

2. Datos de partida

Se utilizará el dataset `prestamos.csv`, que contiene información histórica sobre solicitudes de préstamo, incluyendo variables como:

- Edad del solicitante
- Ingresos mensuales
- Importe del préstamo solicitado
- Duración del préstamo
- Tipo de contrato o situación laboral (si está disponible)
- Variable objetivo: si el préstamo fue impagado o no

Este dataset simula adecuadamente las condiciones de una pyme que evalúa solicitudes de financiación con información limitada pero suficiente.

2.1. Alternativas en Kaggle

Además del dataset propio, existen fuentes públicas de datos en Kaggle que pueden utilizarse como base alternativa o de comparación:

- Home Credit Default Risk: https://www.kaggle.com/competitions/home-credit-default-risk
  Dataset muy completo, ideal para análisis profundos, aunque complejo de procesar.
- Loan Default Dataset: https://www.kaggle.com/datasets/yasserh/loan-default-dataset
  Sencillo y adecuado para un MVP, con variables básicas como ingresos, edad e impago.
- Credit Risk Dataset: https://www.kaggle.com/datasets/laotse/credit-risk-dataset
  Bien estructurado, incluye tasas de interés, ingresos y estado del préstamo.

Estas fuentes permiten ampliar el análisis, probar modelos o comparar resultados con distintos conjuntos de datos.

3. Modelado de la probabilidad de impago (PD)

3.1. Selección de variables

Se recomienda usar solo variables que una pyme pueda recopilar de forma sencilla al recibir una solicitud:

Variable	¿Relevante?	¿Justificación?
Edad	Sí	Factor indirecto de estabilidad y riesgo
Ingresos	Sí	Directamente ligado a la capacidad de pago
Importe solicitado	Sí	Cuanto más alto, mayor riesgo
Duración del préstamo	Sí	Relacionado con exposición a incertidumbre
Tipo de contrato	Sí	Puede afectar a la estabilidad financiera

3.2. Modelos recomendados

Es importante destacar que, aunque la regresión logística se utiliza para predecir una variable categórica (por ejemplo, impago sí/no), el modelo genera como salida una **probabilidad** de que el resultado sea "sí" (impago). Esa probabilidad, que toma valores continuos entre 0 y 1, es lo que se denomina **probabilidad de default (PD)**. Por tanto, aunque el modelo no predice una variable continua directamente, sí proporciona una estimación continua de riesgo que se puede utilizar como base para decisiones de negocio como el cálculo de pricing ajustado al riesgo.

Este mismo principio aplica a otros modelos de clasificación como el **árbol de decisión** o el **Random Forest**: además de predecir la clase (impago/no impago), pueden calcular la **probabilidad de pertenencia a la clase impago**. En el caso del árbol de decisión, esta probabilidad se estima como la proporción de ejemplos positivos en la hoja donde cae la observación. En Random Forest, se obtiene como el promedio de las

probabilidades predichas por cada árbol del conjunto. Así, también permiten estimar una PD continua útil para personalizar el tipo de interés según el riesgo individual del cliente.

Modelo	Ventajas	Inconvenientes
Regresión logística	Simple, interpretable, rápida de entrenar. Ideal para explicar el modelo a no técnicos.	Solo capta relaciones lineales; menos potente con interacciones complejas.
Árbol de decisión	Capta relaciones no lineales, fácil de visualizar y explicar.	Puede sobreajustarse si no se poda o regula bien.
Random Forest (opcional)	Mayor precisión y estabilidad que un único árbol.	Más complejo, difícil de explicar a usuarios no técnicos. Menos recomendable si se prioriza simplicidad.

Para un MVP en 4 semanas, la mejor combinación es:

1. Regresión logística como baseline
2. Validar con árbol de decisión simple para explorar mejoras

3.3. Métrica de evaluación: AUC

La calidad del modelo se evaluará con AUC (Area Under the Curve), que mide la capacidad del modelo para distinguir entre clientes que impagan y los que no. Un AUC de 0.5 equivale a adivinar al azar; un AUC superior a 0.8 ya indica un buen rendimiento.

4. Cálculo del interés mínimo ajustado al riesgo

Una vez entrenado el modelo y obtenida la probabilidad de default (PD) para cada cliente, se calculará un tipo de interés mínimo recomendado para cubrir el riesgo esperado. Esto se puede hacer mediante una fórmula simple, como:

Interés mínimo = tipo base + (coeficiente de riesgo * PD)

- **Tipo base**: un valor fijo que represente el coste de oportunidad o coste financiero mínimo (ej. 5%).
- **Coeficiente de riesgo**: margen adicional proporcional al riesgo (ej. 10% o 15%).
- **PD**: salida del modelo predictivo (valor entre 0 y 1).

Ejemplo:
Un cliente con PD = 0.30 y tipo base = 5%, coeficiente = 10%, tendría:
Interés mínimo = 5% + (0.30 * 10%) = 8%

5. MVP final esperado

En 4 semanas, el proyecto puede entregar:

- Modelo entrenado de predicción de PD
- Evaluación del modelo con métricas AUC y visualización ROC
- Fórmula funcional de cálculo de interés mínimo
- Tabla o app interactiva (por ejemplo, en Streamlit) para simular distintos perfiles de cliente
- Informe explicativo del modelo y justificación del pricing

6. Posibles extensiones futuras

- Añadir detección de anomalías para solicitudes sospechosas
- Introducir criterios éticos para evitar sesgos en la concesión
- Conectar la herramienta con formularios reales de solicitud (API o Excel)

Este enfoque permite al alumno desarrollar una solución realista, útil y diferenciadora, aplicando técnicas de machine learning en un contexto empresarial accesible. Además, sienta las bases para futuras ampliaciones más complejas si se desea continuar tras el MVP.