Modelos de Machine Learning para la Predicción de Impago en el Sector Financiero

Máster en Big Data & Business Analytics
Maria Regina Meiners de Alba, Eva Meneses Soto y Ana María Quintero Estévez
Madrid, 2025

Roadmap

01

Default bancario

Interés de estudio y objetivos

02

Dataset

Descripción y variables

03

Metodología

Pipeline de análisis, modelado y visualización

04

EDA y Preprocesamiento

Análisis Exploratorio de Datos
Limpieza, transformación y preparación para el modelado

05

Modelos

Proceso y selección de métricas

06

Evaluación e interpretabilidad

Comparación de métricas y selección del modelo

07

Dashboards

Seguimiento de KPIs y monitoreo del modelo

08

Conclusiones

Principales hallazgos y aportes del estudio

01

Default bancario

Interés de estudio y objetivos

¿Por qué predecir el impago?

- El default bancario afecta la salud financiera del sector y la economía.
- Predecir impagos a tiempo ayuda a reducir riesgos y optimizar decisiones.
- Necesidad de modelos explicables y precisos para la gestión crediticia.

Objetivo

Optimizar la gestión del riesgo crediticio mediante modelos de Machine Learning que anticipen e identifiquen el impago de forma más eficiente que los métodos tradicionales.

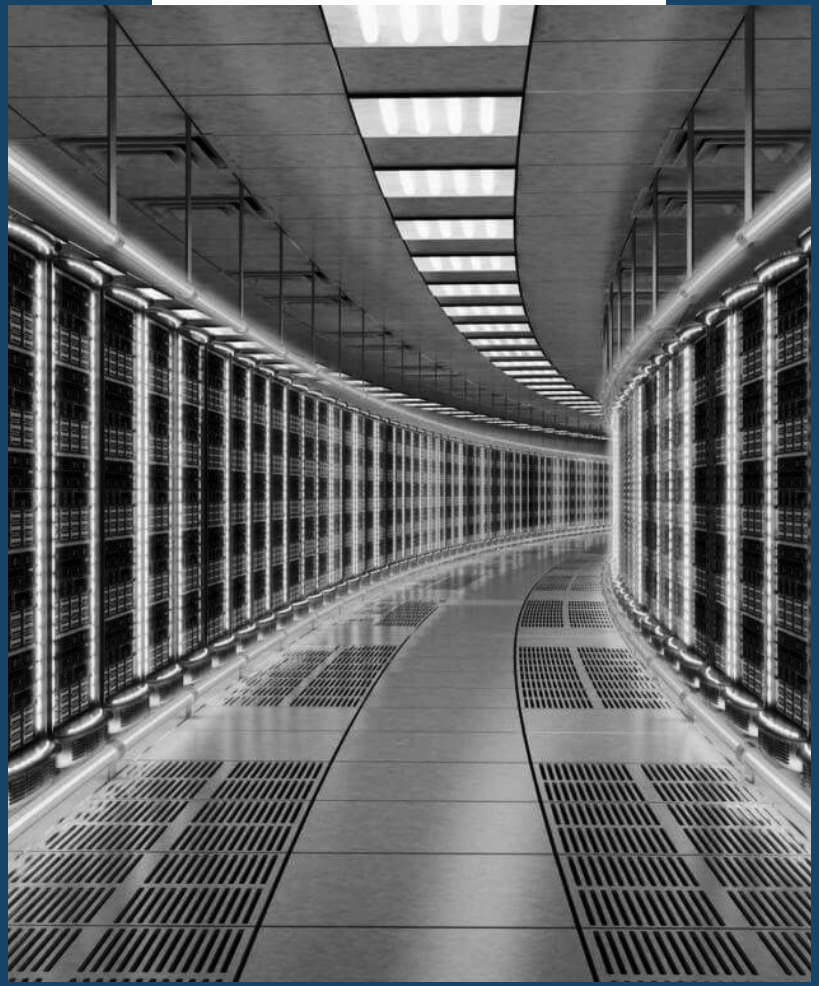
Nos centraremos en tres partes principales:

- Comparación de modelos predictivos lineales y no lineales
- Interpretabilidad de los modelos
- Dashboards de seguimiento y monitoreo

02

Dataset

Características y variables



Dataset

Loan Dataset

MIT

Descripción

- Datos demográficos de los clientes
- Detalles del préstamo

+32.000 Registros

13 Variables

Variables

1

Variable objetivo
Estado del préstamo

8

Numéricas
Id, edad, ingresos, importe, tasa de interés, antigüedad laboral, plazo y longitud del historial crediticio.

4

Categóricas
Tipo de vivienda, intención del crédito, calificación e historial de impago.

03

Metodología

Pipeline de análisis, modelado y visualización

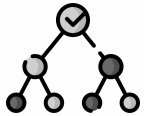
Metodología

EDA

- Identificación valores erróneos
- Estadísticas descriptivas y visualizaciones iniciales

Preprocesamiento

- Tratamiento de valores nulos y erróneos
- Recodificación de variables



Modelado

- Entrenamiento de modelos lineales y no lineales
- Selección y ajuste de hiperparámetros

Evaluación

- Comparativa de modelos (Fmeas)
- Interpretabilidad (SHAP, análisis de variables clave)

Visualización

- Creación de dashboards de seguimiento y monitoreo

Mejora contínua

- Monitoreo de métricas clave en el dashboard
- Revisión periódica del modelo
- Propuesta de re-entrenamiento ante cambios en los datos

04

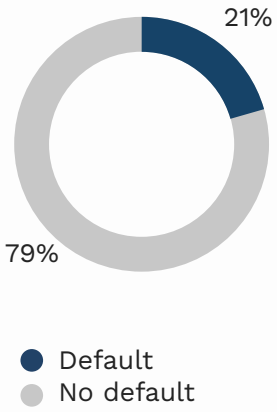
EDA y Preprocesamiento

Análisis Exploratorio de Datos
Limpieza, transformación y preparación para el modelado

Análisis Exploratorio de Datos

Detección de prevalencia

Estado del crédito



Valores nulos

Historial de impago

63,6%

Tasa de interés

9,57%

Antigüedad laboral

2,75%

Valores erróneos

Edad

<18 años
>100 años

Antigüedad laboral

> 45 años

Selección de variables

p-valor

< 0,05

**Sin customer_id*

Inclusión de todas las variables en el modelo:

- Asociación significativa con el objetivo
- Valor predictivo

Preprocesamiento

Limpieza de errores

Tratamiento de valores incoherentes para asegurar la calidad del dataset

Eliminación de registros

Registros sin antigüedad laboral o duplicados

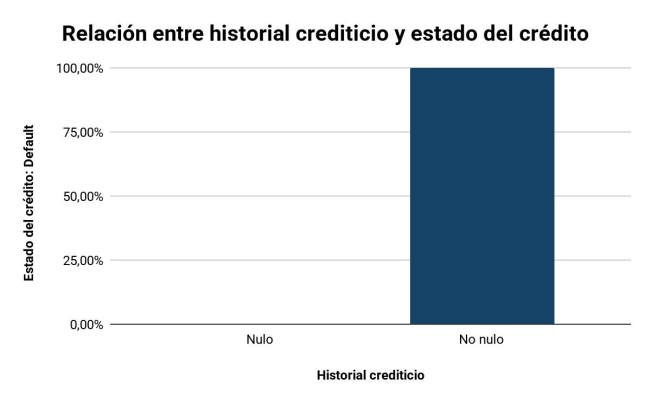
Recodificación de variables

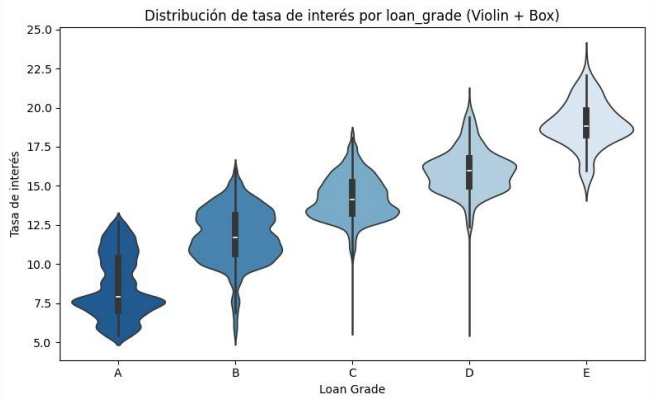
Conversión de variables como home_ownership, loan_intent, y loan_grade a string

Tratamiento de valores nulos

Elección de métodos de imputación adecuados para cada caso

Tratamiento de valores nulos





Historial de impago

Reemplazo de valores nulos por “No register”

Tasa de interés

Tratamiento en receta por mediana del grupo

05

Modelos

Proceso y selección de métricas

Modelos lineales

Ridge

- Capacidad para reducir el sobreajuste
- Manejar bien la multicolinealidad
- Manteniendo los predictores sin eliminar información relevante

Lasso

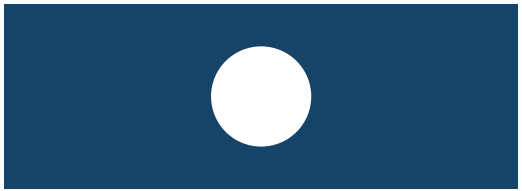
- Reducir el sobreajuste
- Selección de variables al eliminar las irrelevantes
- Mejora la interpretabilidad

Puede perder información útil

Elastic net

- Combina las ventajas de Ridge y Lasso
- Equilibra la reducción de sobreajuste y selección de variables

Solo se prefiere si mejora significativamente



Modelos no lineales

Árboles de decisión	KNN	SVM	MARS
Bagging Se considera útil si todas las variables son importantes **Random Forest** Ofrece un buen equilibrio entre sesgo y varianza	Identificar si clientes con características similares han hecho default: • Agrupa • Clasifica nuevos registros según la distancia entre datos	Es útil cuando hay una clara separación entre clases Encuentra el hiperplano óptimo para clasificar entre quienes hacen y no hacen default	Combina la simplicidad de los modelos lineales con la flexibilidad de los no lineales Modela relaciones complejas mediante funciones por tramos (splines)

Proceso

Splitting

Estratificado:
70% Training
30% Testing

Preprocesamiento

Recetas adaptadas a las particularidades de cada modelo

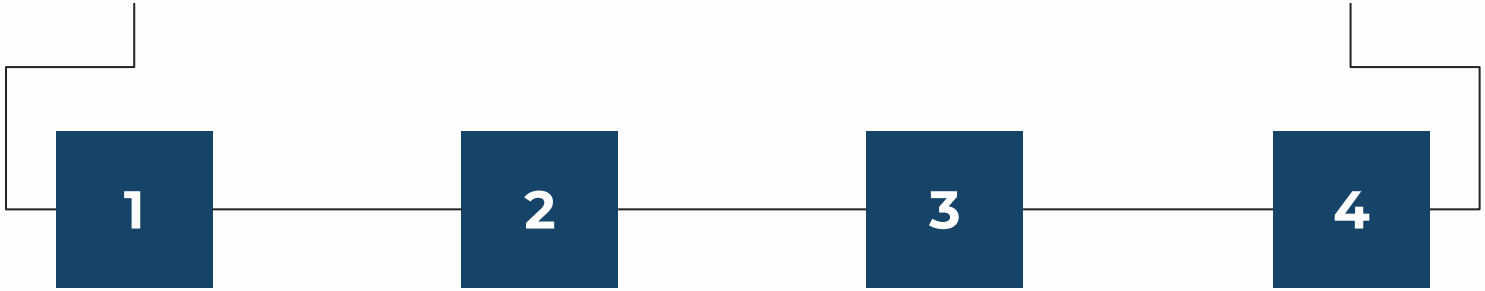
Validación cruzada

2 conjuntos de datos:

- 5 particiones
- 2 repeticiones
- Diferentes semillas

Selección de hiperparámetros

Grid: max entropy
Expand grid



Selección de la métrica

Métrica: f-meas

Concepto: mide el equilibrio entre sensibilidad y precisión:

- Capacidad del modelo para identificar correctamente los positivos
- Proporción de positivos predichos que realmente lo son

Objetivo: equilibrar precisión y sensibilidad en contextos de riesgo

06

Evaluación e interpretabilidad

Comparación de métricas y selección del modelo

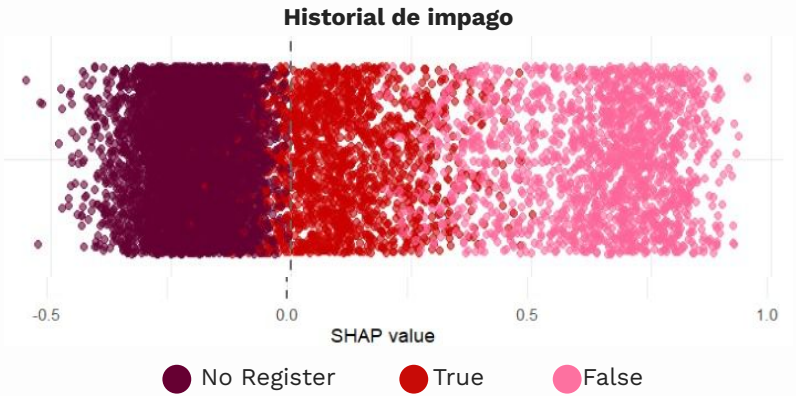
Comparación de métricas

	Lasso	Random forest
F-Measure	0,830	0,929
Especificidad	0,956	0,991
Sensibilidad	0,828	0,895
Curva ROC	0,978	0,993
Precisión	0,833	0,965
Detección prevalencia	0,207	0,193
Acuracidad	0,930	0,971

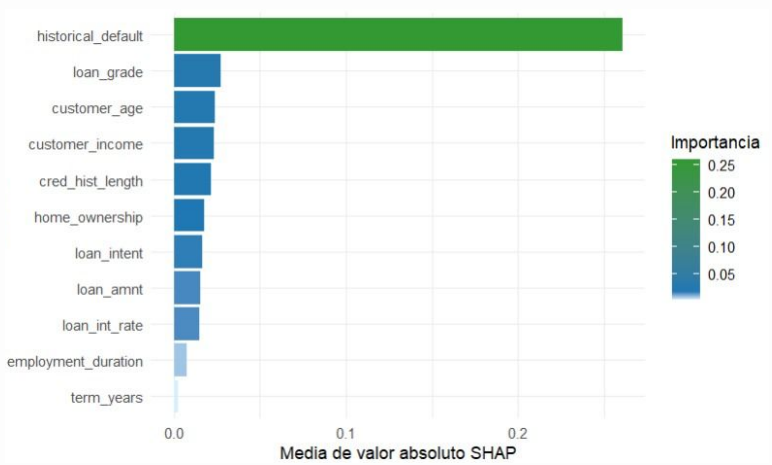
Interpretabilidad

SHAP values

- Contribución que cada una de las variables en la predicción
- Visión transparente del proceso de decisión del modelo



Importancia global de variables



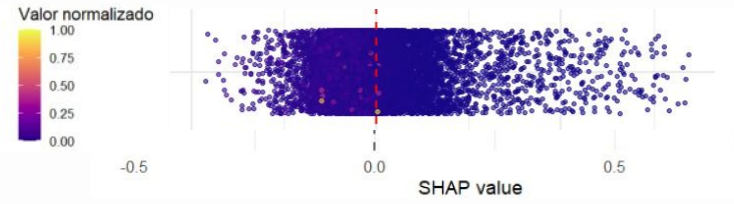
Comparación de métricas

*Sin historial de impago	Lasso	Random forest
F-Measure	0,568	0,806
Especificidad	0,952	0,986
Sensibilidad	0,469	0,712
Curva ROC	0,853	0,933
Precisión	0,720	0,929
Detección prevalencia	0,135	0,159
Acuracidad	0,852	0,929

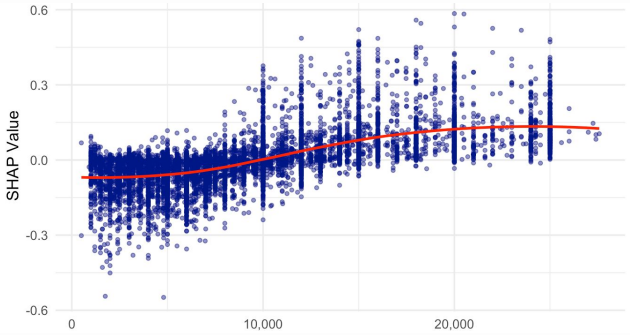
Interpretabilidad

SHAP values: Numéricas

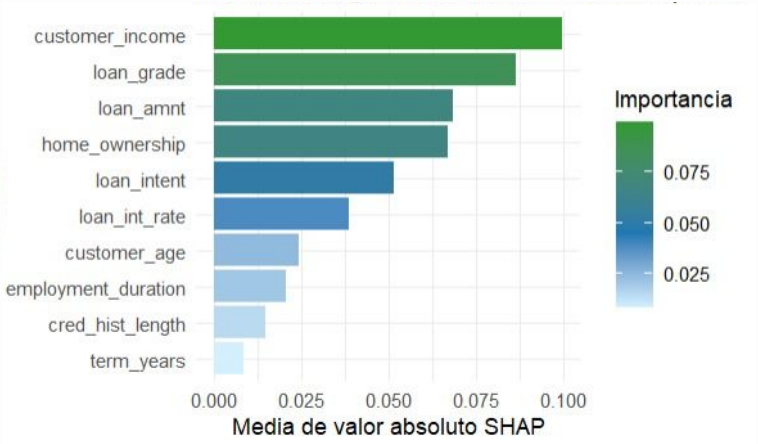
Ingreso del cliente



Monto del crédito

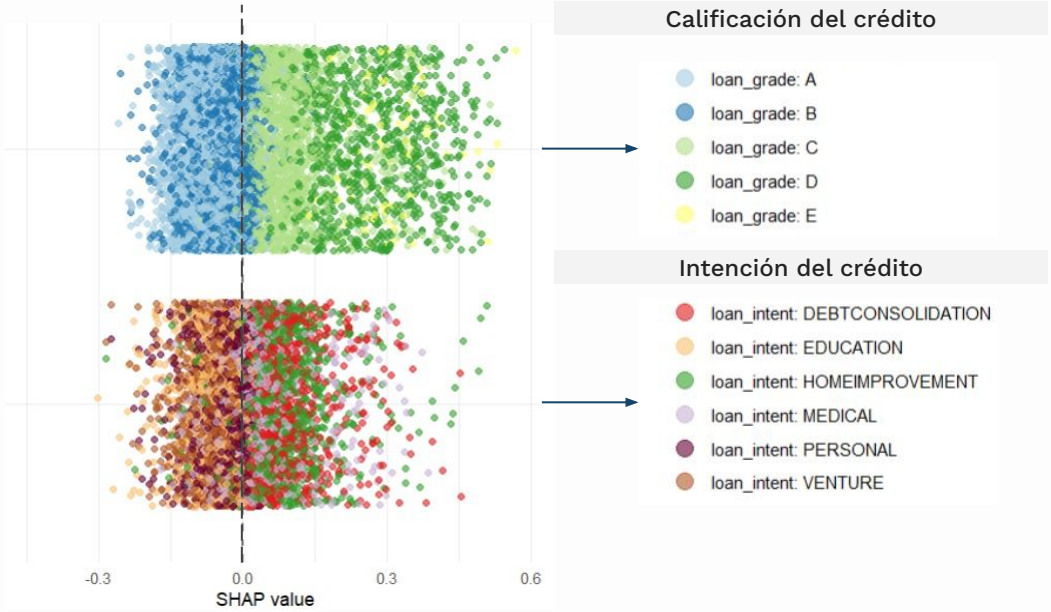


Importancia global de variables



Interpretabilidad

SHAP values: Categóricas



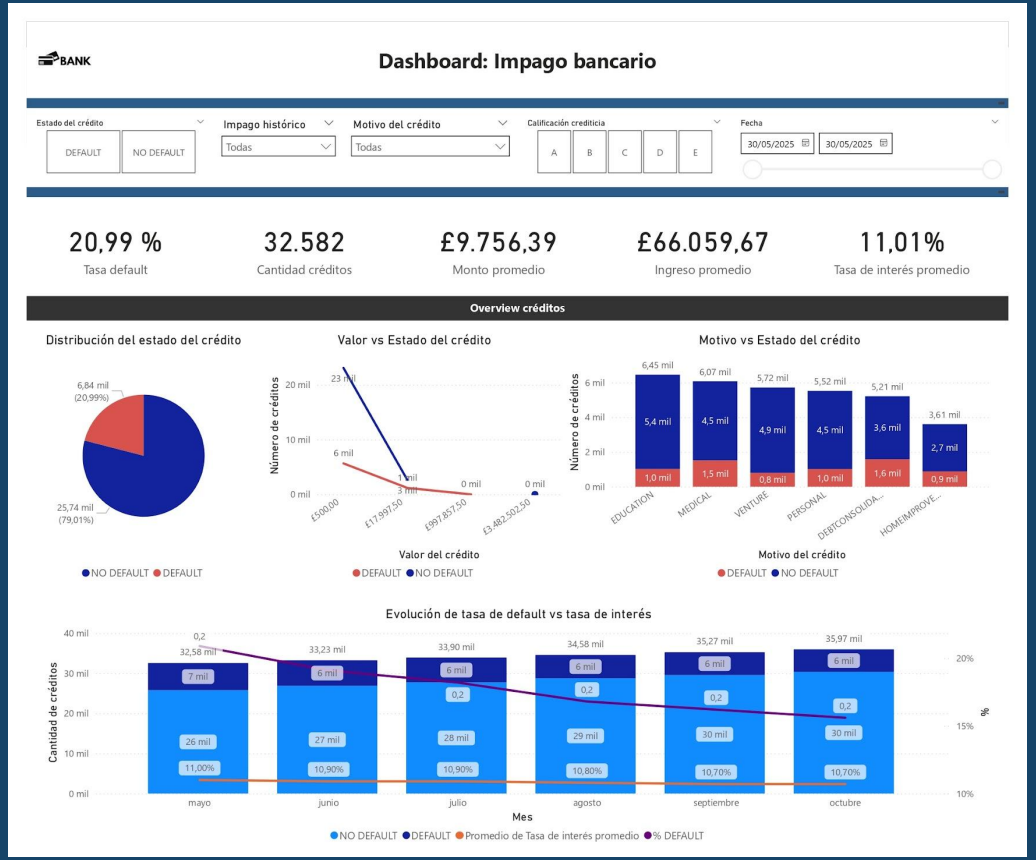
07

Dashboards

Seguimiento de KPIs y monitoreo del modelo

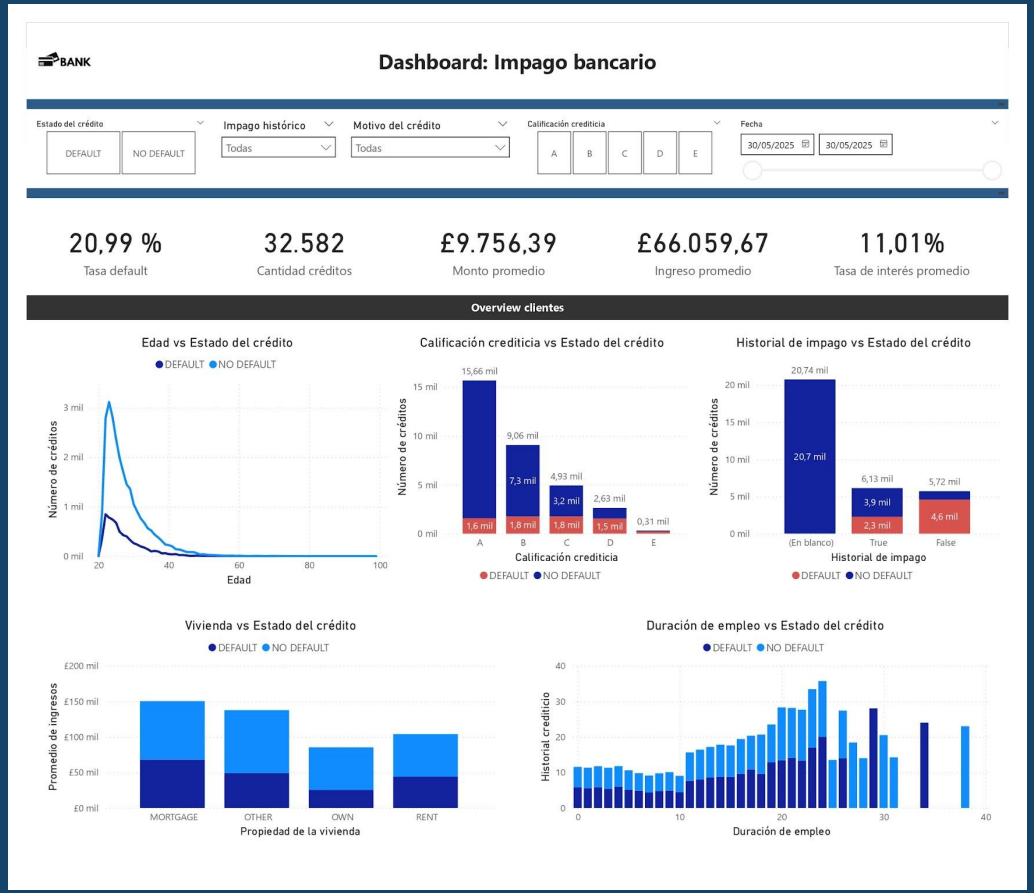
Dashboard de negocio

Overview de créditos



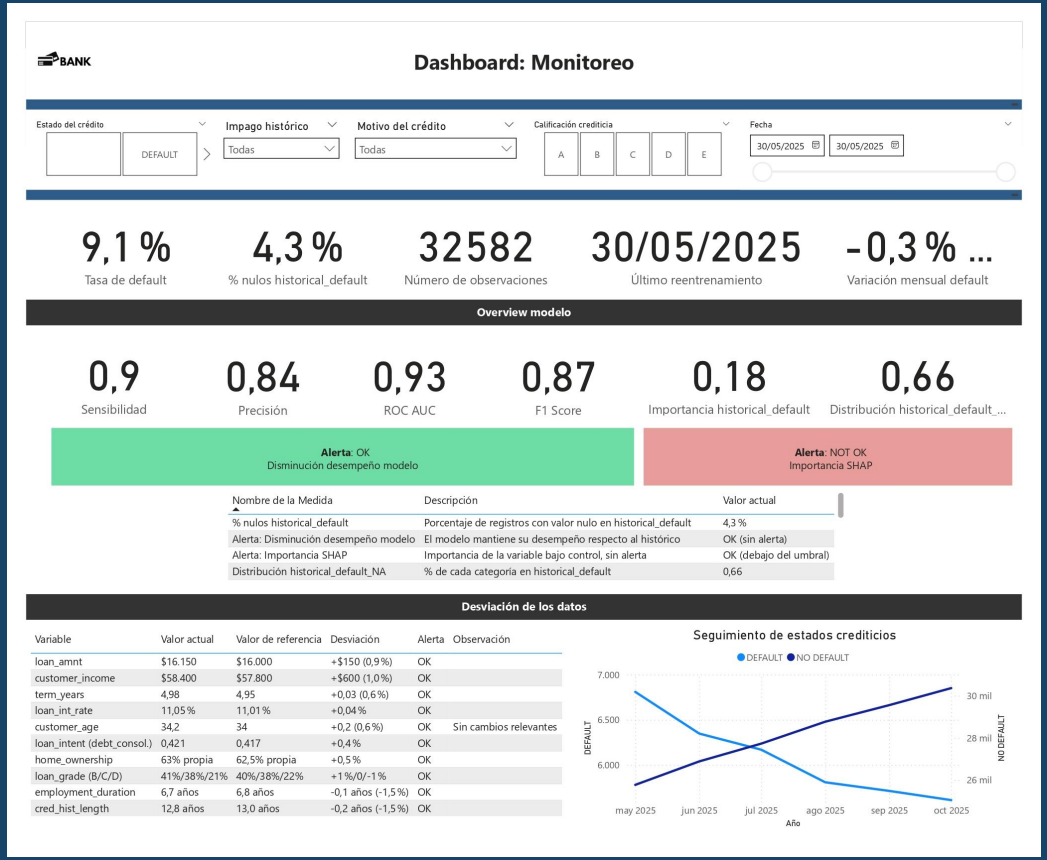
Dashboard de negocio

Overview de clientes



Dashboard de monitoreo

Overview del modelo y el dataset



08

Conclusiones

Principales hallazgos y aportes del estudio

Conclusiones

01

Modelos

Los modelos no lineales se adaptan mejor en este escenario

02

Variables clave

Ingresos, calificación crediticia e historial de impago

03

Hallazgo relevante

Valores nulos no aleatorios en historial de impago

04

Transparencia

Priorizamos evitar sesgos sobre la precisión del modelo

05

Seguimiento de KPIs

Traducción de desempeño en indicadores de negocio

06

Monitoreo del modelo

Observar el comportamiento del modelo y cambios en el dataset

Limitaciones

- Uso de un dataset público
- Ausencia de variables externas reducen el alcance predictivo
- Almacenamiento

Líneas futuras

- Incluir datos macroeconómicos o comportamentales
- Evaluar el impacto en indicadores de negocio

¡Gracias!

Espacio de preguntas