

ANÁLISIS ENTIDAD CREDITICIA



LUISA FERNANDA SUAZO CASTAÑO

DIANA MARCELA ROMERO CASTAÑO

INTRODUCCIÓN



- En la actualidad, el análisis del endeudamiento crediticio se ha convertido en una herramienta fundamental para entidades financieras, aseguradoras, fintechs y otras organizaciones interesadas en comprender el comportamiento financiero de sus clientes. A través del uso de técnicas de analítica de datos, es posible identificar factores que inciden en el nivel de deuda de una persona, permitiendo desarrollar modelos predictivos que anticipen comportamientos financieros de riesgo, fomenten decisiones de crédito más responsables y promuevan la educación financiera



OBJETIVO

- Desarrollar un análisis integral del comportamiento crediticio de los clientes a partir de un conjunto de datos financieros, aplicando técnicas de analítica descriptiva y predictiva que permitan identificar patrones relevantes, explorar relaciones entre variables y construir modelos estadísticos para explicar y predecir fenómenos asociados al endeudamiento y riesgo de incumplimiento



CONJUNTO DE DATOS:

Explicación de la característica:

- **D:** Identificador único del cliente.
- **Default:** Si el cliente tiene más de 90 días sin pagar su préstamo (1 = Sí, 0 = No).
- **Prct_uso_tc:** Porcentaje de uso de tarjeta de crédito.
- **Edad:** Edad del cliente.
- **Nro_prestao_retrasados:** Número de préstamos con retraso en el pago.
- **Prct_deuda_vs_ingresos:** Porcentaje de deuda con respecto a los ingresos.
- **Mto_ingreso_mensual:** Monto de ingreso mensual.
- **Nro_prod_financieros_deuda:** Número de productos financieros con deuda.
- **Nro_retraso_60dias:** Número de retrasos en pagos mayores a 60 días.
- **Nro_creditos_hipotecarios:** Número de créditos hipotecarios.
- **Nro_retraso_ult3anios:** Número de retrasos en los últimos 3 años.
- **Nro_dependiente:** Número de personas dependientes del cliente.

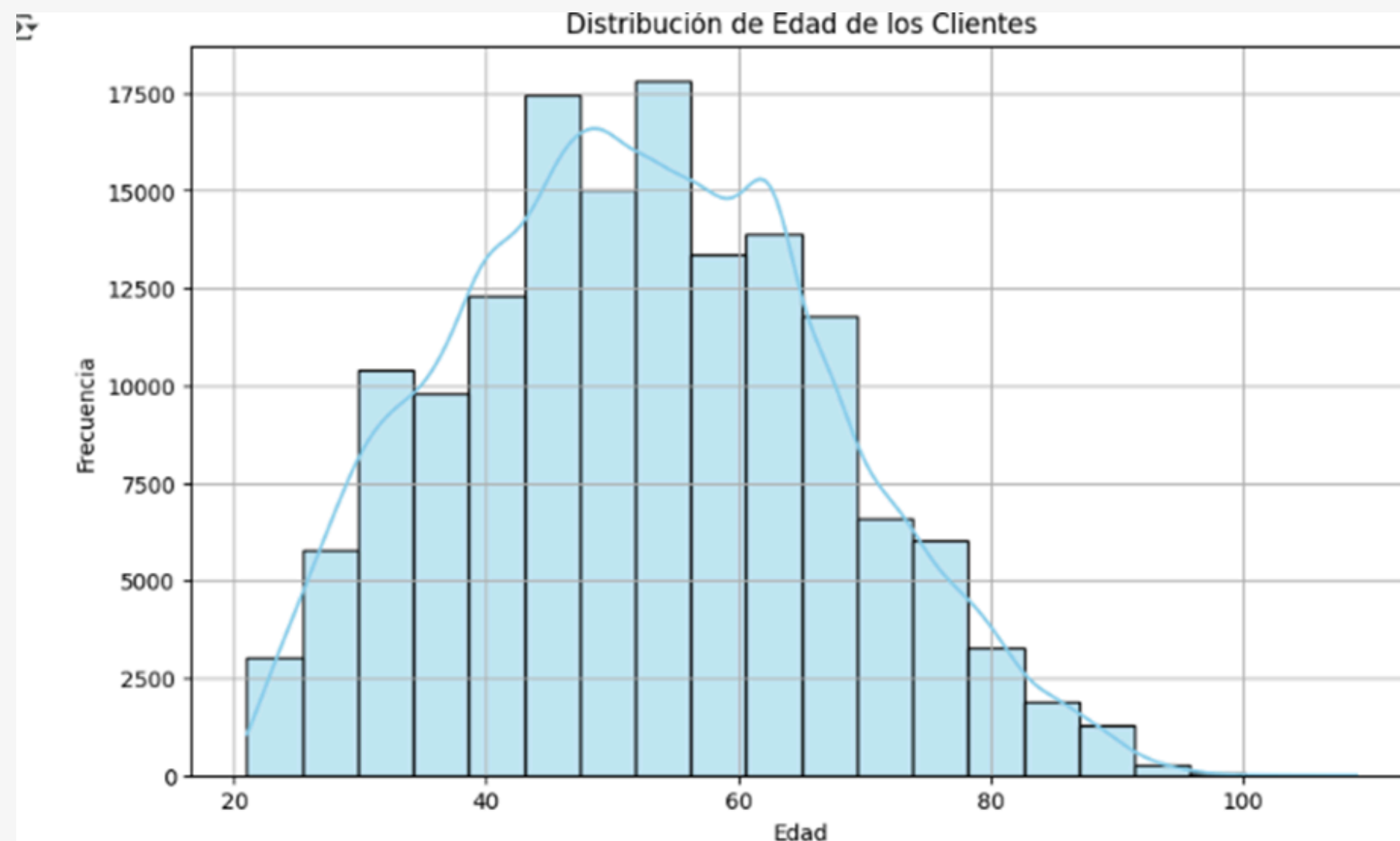
01



ANÁLISIS EXPLORATORIO DE DATOS (EDA):

- Responder preguntas descriptivas usando estadísticas y visualizaciones

Pregunta Descriptiva 1: ¿Cuál es la distribución de edades de los clientes?

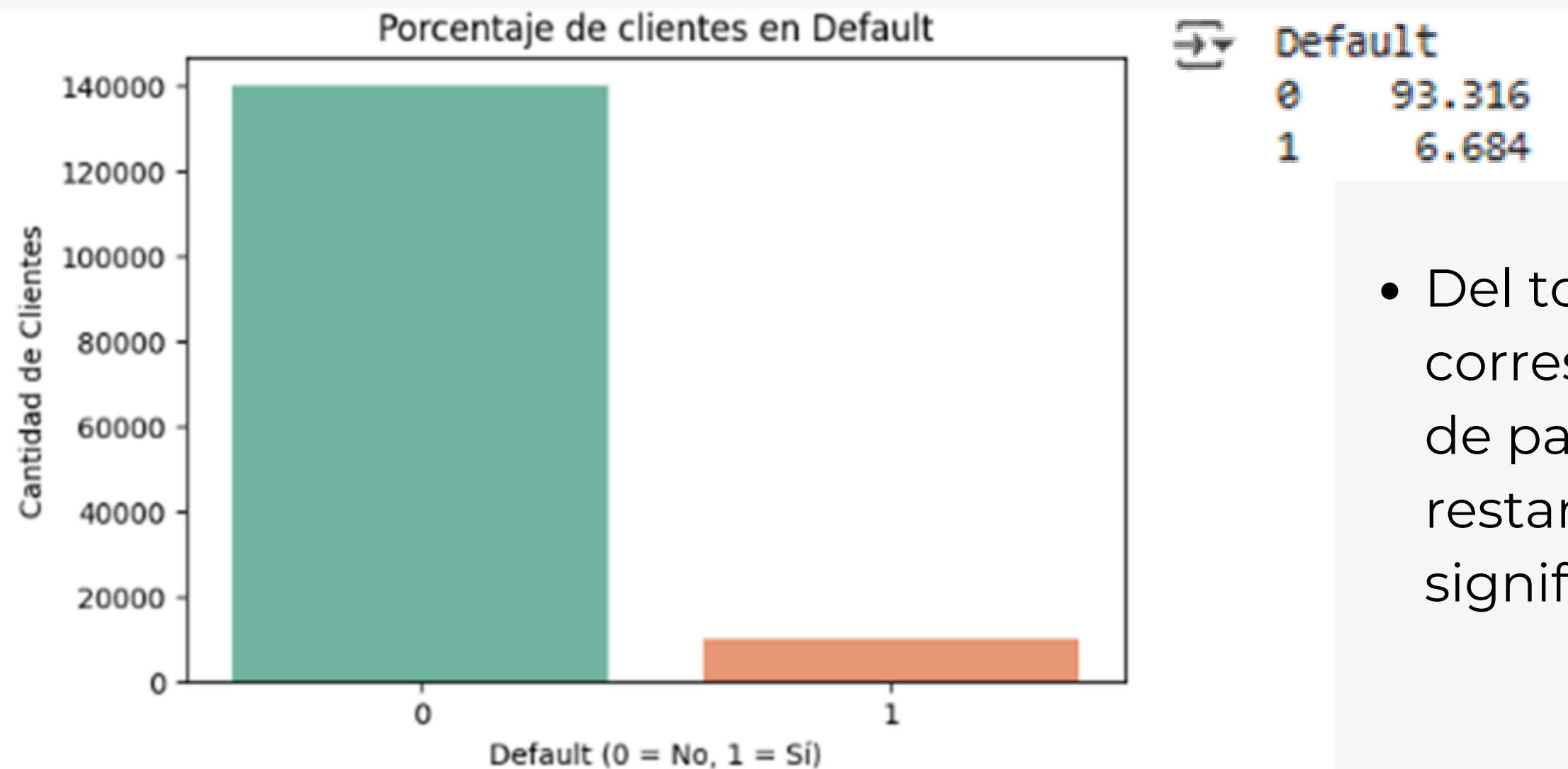


- La variable Edad presenta una distribución centrada y simétrica, con una media y mediana cercanas a los 52 años, lo que indica que la mayoría de los clientes son adultos en edad laboral.
- El rango intercuartílico (IQR) indica que el 50% de los clientes tienen entre 41 y 63 años, lo cual es típico en carteras activas de crédito.
- El valor mínimo ahora es 21 años, lo cual es coherente con la edad mínima esperada para tener acceso a productos financieros.

ANÁLISIS EXPLORATORIO DE DATOS (EDA):

- Responder preguntas descriptivas usando estadísticas y visualizaciones

Pregunta Descriptiva 2: ¿Qué porcentaje de los clientes ha caído en default (Default = 1)?

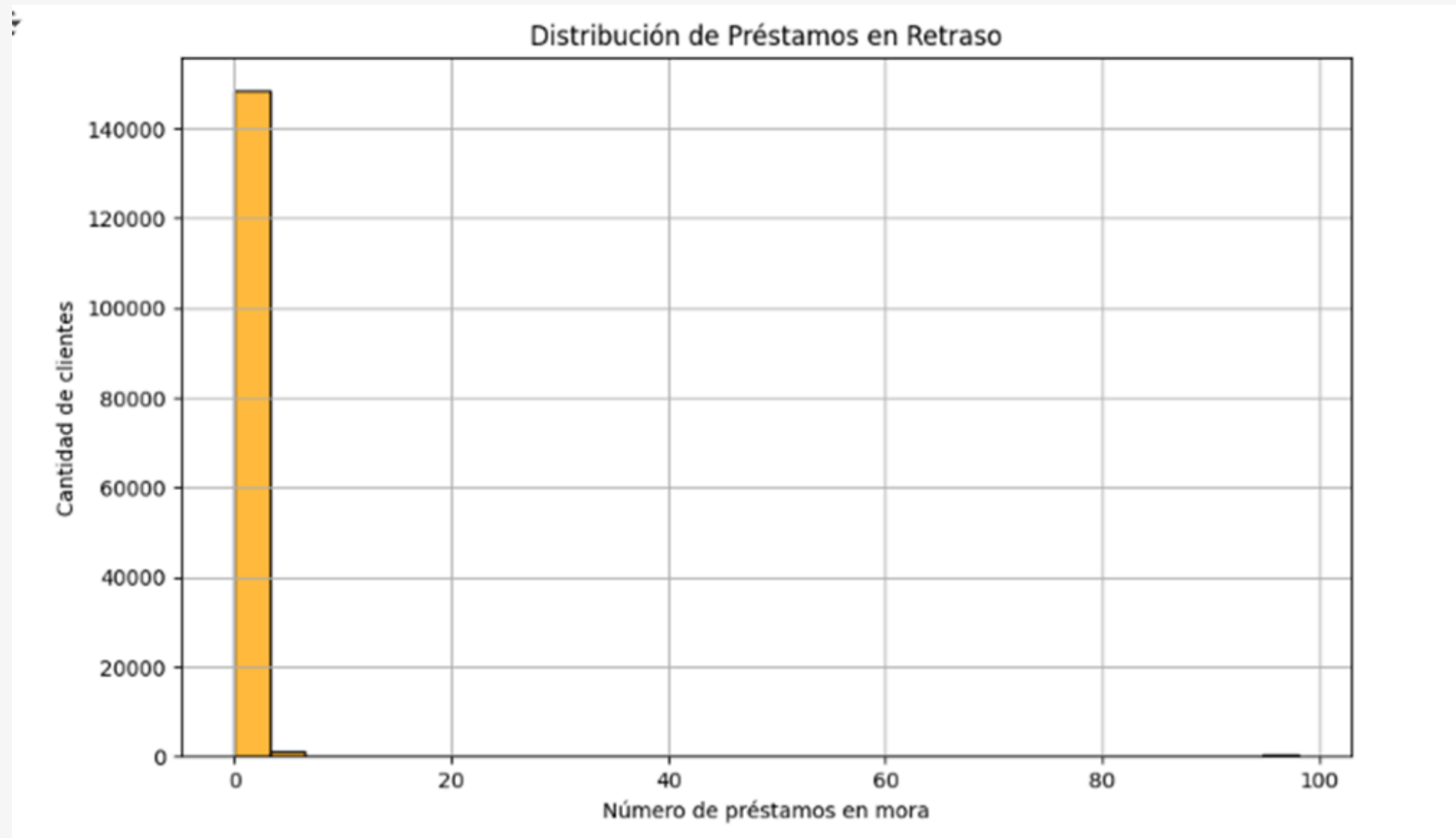


- Del total de registros analizados, el 6.68% corresponde a clientes en incumplimiento de pago (Default = 1), mientras que el restante 93.32% no presenta mora significativa.

ANÁLISIS EXPLORATORIO DE DATOS (EDA):

- Responder preguntas descriptivas usando estadísticas y visualizaciones

Pregunta Descriptiva 3: ¿Cómo se distribuye el número de préstamos en mora?

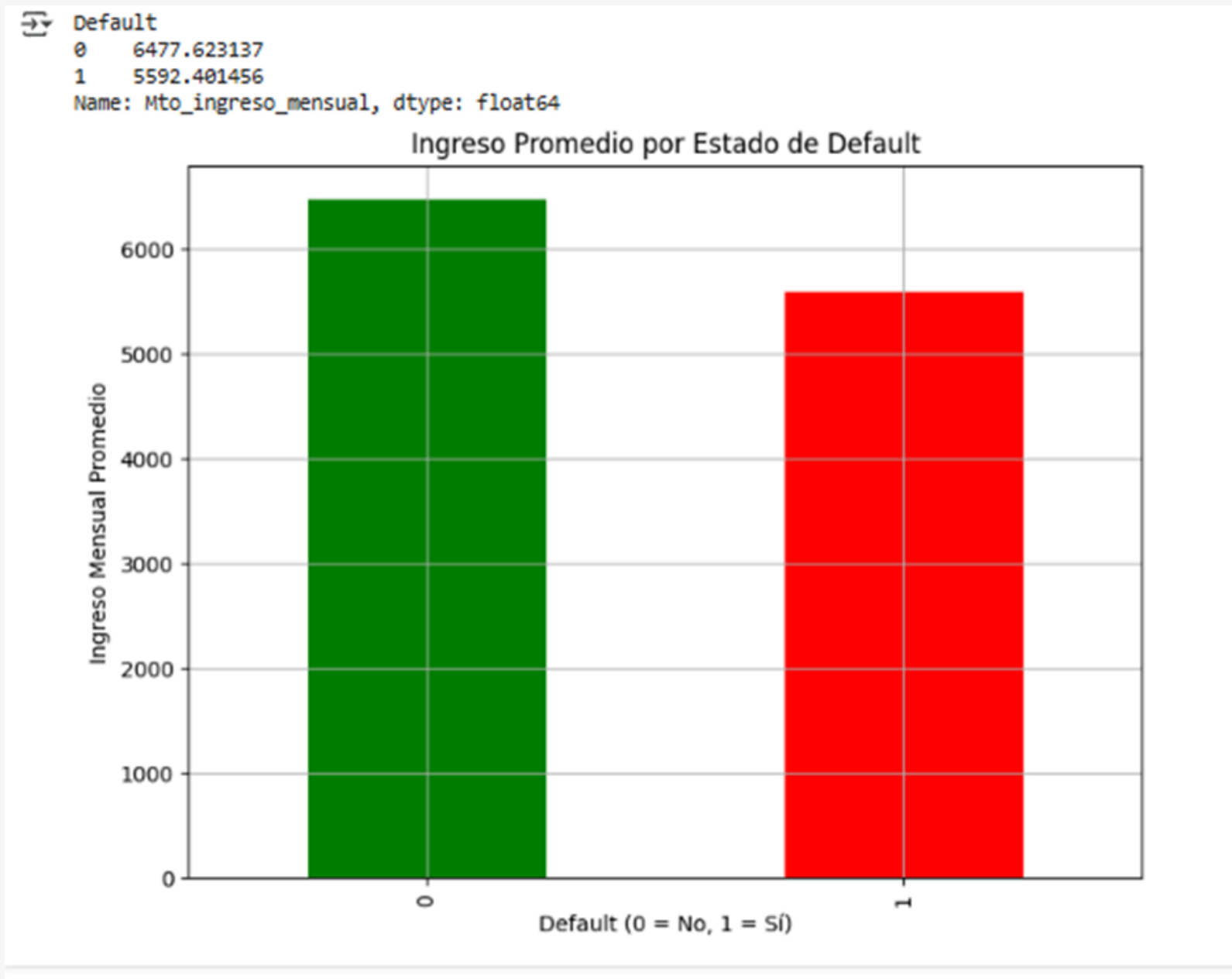


- La mayoría de los clientes no presentan préstamos en mora. Más de 120,000 registros tienen cero préstamos retrasados, lo que representa una gran proporción de la muestra.

ANÁLISIS EXPLORATORIO DE DATOS (EDA):

- Responder preguntas descriptivas usando estadísticas y visualizaciones

•Pregunta Descriptiva 4: ¿Hay diferencia de ingresos entre clientes en default y los que no?



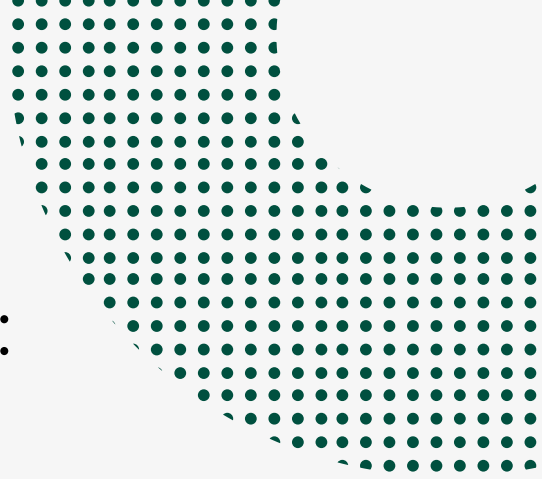
- Los clientes que no han caído en default tienen un ingreso mensual promedio de aproximadamente \$6,477, mientras que los que sí han caído en default tienen ingresos promedio de \$5,592.
- Esto sugiere que el nivel de ingresos podría ser un factor asociado al riesgo de incumplimiento. Además, se observaron valores máximos muy altos en clientes sin default, lo que indica alta dispersión

ANÁLISIS EXPLORATORIO DE DATOS (EDA):

- **Outliers detectados:** Edad máxima de 109 años → se consideró como posible outlier (valor superior a 96).
- **Distribución de clases** (Default): Desbalance fuerte → 93% clase 0 (no default), 7% clase 1 (default).
- Limpieza y transformación e imputación de datos
- **Se aplicó SMOTE** para balancear las clases antes de modelos clasificadores.
- **Variables consideradas:** Edad, ingresos, número de dependientes, porcentaje de deuda vs ingreso, entre otras.



ANÁLISIS EXPLORATORIO DE DATOS (EDA):



Medidas de Dispersión y Simetría

Utilizaremos las variables principales del análisis:

	mean	std	min	max	IQR	Asimetría
Edad	52.295207	14.771866	0.0	1.090000e+02	22.000000	0.188993
Mto_ingreso_mensual	6418.454920	12890.395542	0.0	3.008750e+06	3497.000000	127.120424
Prct_uso_tc	230.488315	315.173185	0.0	8.851852e+03	343.014259	2.437326
Prct_deuda_vs_ingresos	265.887465	417.869993	0.0	9.906298e+03	382.597738	7.101164

Las medidas de dispersión mostraron que el ingreso mensual (**Mto_ingreso_mensual**) es una de las variables con mayor variabilidad, con una desviación estándar considerablemente alta respecto a su media. Esto sugiere que hay clientes con ingresos muy dispares, probablemente debido a valores atípicos.

La variable Edad tiene una dispersión mucho menor, con la mayoría de clientes en el rango de 30 a 70 años.

En cuanto a la asimetría (skewness):

Prct_uso_tc y **Prct_deuda_vs_ingresos** presentan asimetría positiva, indicando una distribución sesgada hacia la derecha, con la mayoría de los valores concentrados en niveles bajos pero algunos clientes con porcentajes extremadamente altos.

Mto_ingreso_mensual también presenta alta asimetría, confirmando la presencia de clientes con ingresos excepcionalmente altos.

MODELOS APLICADOS Y RESULTADOS

REGRESIÓN LINEAL MULTIPLE

Hipótesis Planteada

Hipótesis Nula (H_0):

No existe una relación significativa entre el número de dependientes, el ingreso mensual y la edad del cliente con el porcentaje de deuda frente a sus ingresos.

Hipótesis Alternativa (H_1):

Al menos una de las variables (número de dependientes, ingreso mensual o edad) tiene un efecto significativo sobre el porcentaje de deuda frente a los ingresos del cliente

Objetivo: Entender relación entre variables e indicador de deuda.

Variable dependiente: Prct_deuda_vs_ingresos.

Variable independientes:

Nro_dependiente

Mto_ingreso_mensual

Edad

MODELOS APLICADOS Y RESULTADOS

REGRESIÓN LINEAL MULTIPLE

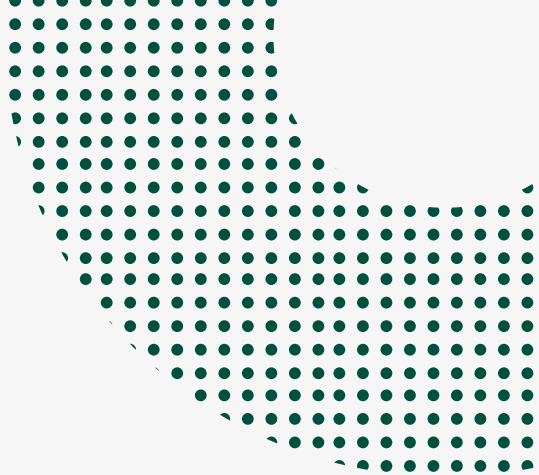
→

OLS Regression Results						
Dep. Variable:	Prct_deuda_vs_ingresos	R-squared:	0.014			
Model:	OLS	Adj. R-squared:	0.014			
Method:	Least Squares	F-statistic:	698.3			
Date:	Thu, 24 Apr 2025	Prob (F-statistic):	0.00			
Time:	17:47:48	Log-Likelihood:	-1.1171e+06			
No. Observations:	150000	AIC:	2.234e+06			
Df Residuals:	149996	BIC:	2.234e+06			
Df Model:	3					
Covariance Type: nonrobust						
	coef	std err	t	P> t	[0.025	0.975]
const	312.4564	4.255	73.434	0.000	304.117	320.796
Nro_dependiente	35.4525	0.994	35.668	0.000	33.504	37.401
Mto_ingreso_mensual	-0.0014	8.34e-05	-16.260	0.000	-0.002	-0.001
Edad	-1.2240	0.074	-16.458	0.000	-1.370	-1.078
Omnibus:	196018.186	Durbin-Watson:	1.993			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	61734783.756			
Skew:	7.217	Prob(JB):	0.00			
Kurtosis:	101.332	Cond. No.	5.74e+04			

Número de dependientes: fue la única con una relación positiva y significativa con la deuda. Es decir, a mayor cantidad de personas a cargo, mayor es el porcentaje de deuda.

Monto de ingreso mensual: presentó una relación ligeramente negativa, lo que sugiere que mayores ingresos se asocian con menores niveles de deuda relativa, aunque su efecto fue muy pequeño.

Edad: también mostró una relación negativa y significativa, indicando que a mayor edad, menor proporción de deuda frente al ingreso.



MODELOS APLICADOS Y RESULTADOS REGRESIÓN LINEAL MULTIPLE RESPUESTA A LA HIPOSTESIS SEGUN LOS RESULTADOS OLS

"Todas las variables en el modelo resultaron estadísticamente significativas ($p < 0.05$), por lo tanto, se rechazan las hipótesis nulas. Esto indica que edad, número de dependientes e ingreso mensual sí tienen relación con el porcentaje de deuda respecto al ingreso, aunque la capacidad explicativa del modelo es baja."

MODELOS APLICADOS Y RESULTADOS

RANDOMFORESTCLASSIFIER

Hipótesis Planteada

Hipótesis nula (H_0):

Las variables financieras y personales no tienen relación significativa con la probabilidad de incumplimiento (Default).

Hipótesis alternativa (H_1):

Existe una relación significativa entre las variables crediticias y personales (edad, uso de crédito, ingresos, préstamos atrasados, etc.) y la probabilidad de que un cliente caiga en incumplimiento (Default = 1).

Objetivo del modelo:

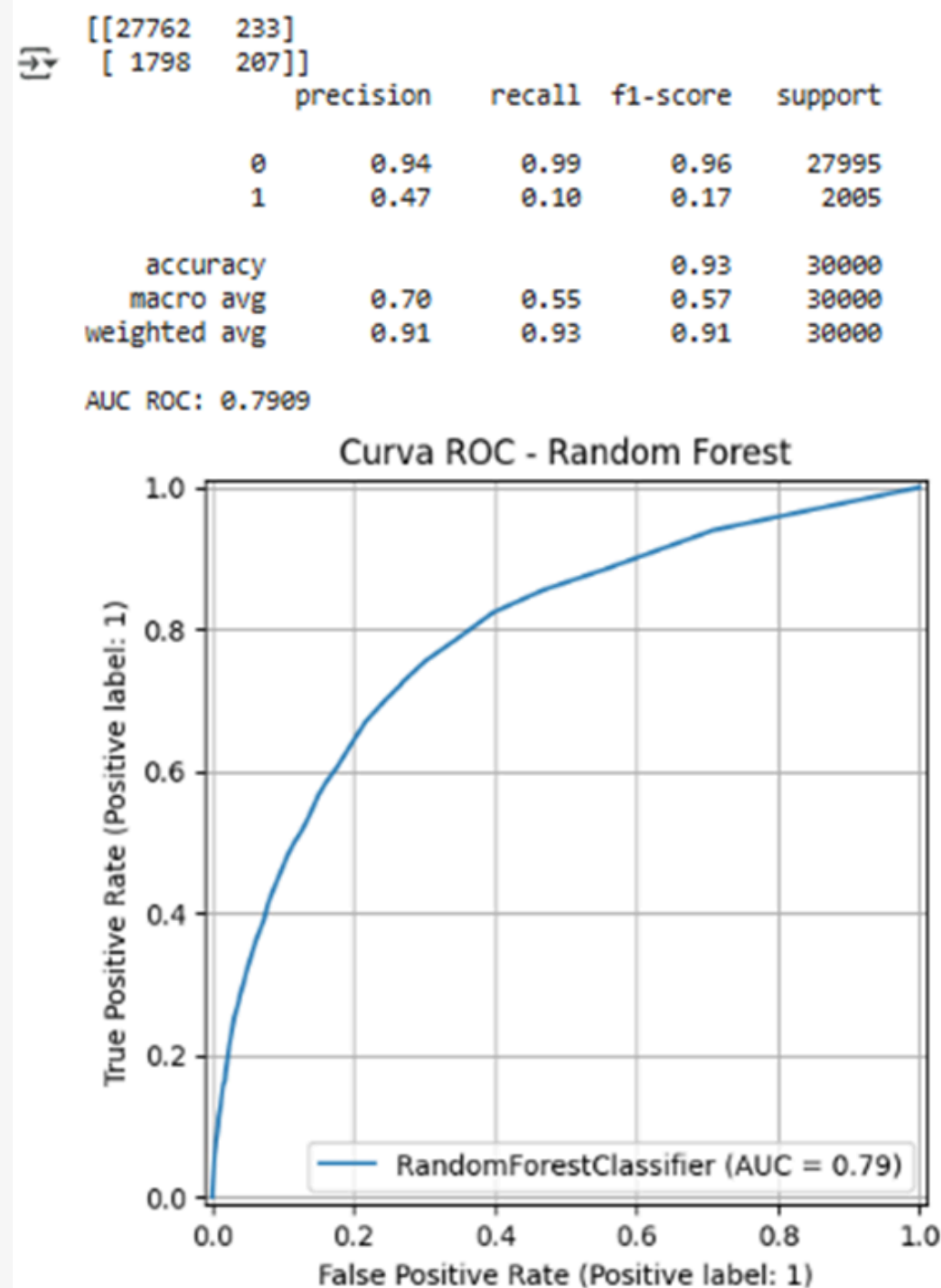
Predecir la probabilidad de que un cliente entre en estado de default (incumplimiento de pago), una variable binaria (0 = no cae en default, 1 = sí cae en default). Es clave para prevenir riesgos crediticios.

Usaremos variables que tienen sentido financiero y que ya limpiaste anteriormente:

- **Edad**
- **Mto_ingreso_mensual**
- **Nro_dependiente**
- **Prct_uso_tc**
- **Nro_prestao_retrasados**
- **Prct_deuda_vs_ingresos**
- **La variable objetivo será: Default**

MODELOS APLICADOS Y RESULTADOS

RANDOMFORESTCLASSIFIER



El modelo predice muy bien la clase 0 (personas que no caen en default), pero mal la clase 1 (personas que sí caen en default).

Esto indica desbalance de clases que afecta el desempeño en la clase minoritaria (1).

El modelo Random Forest mostró una alta capacidad para predecir correctamente a los clientes que no caen en default (clase 0), pero un desempeño deficiente en la detección de quienes sí lo hacen (clase 1). A pesar de una precisión general del 93% y un AUC de 0.79, la capacidad del modelo para detectar riesgo real es limitada (recall de solo 10% en la clase 1).

Por lo tanto, no se rechaza completamente la hipótesis nula, ya que no logra predecir con eficacia el comportamiento de los clientes en default, posiblemente por el desbalance de clases.

MODELOS APLICADOS Y RESULTADOS

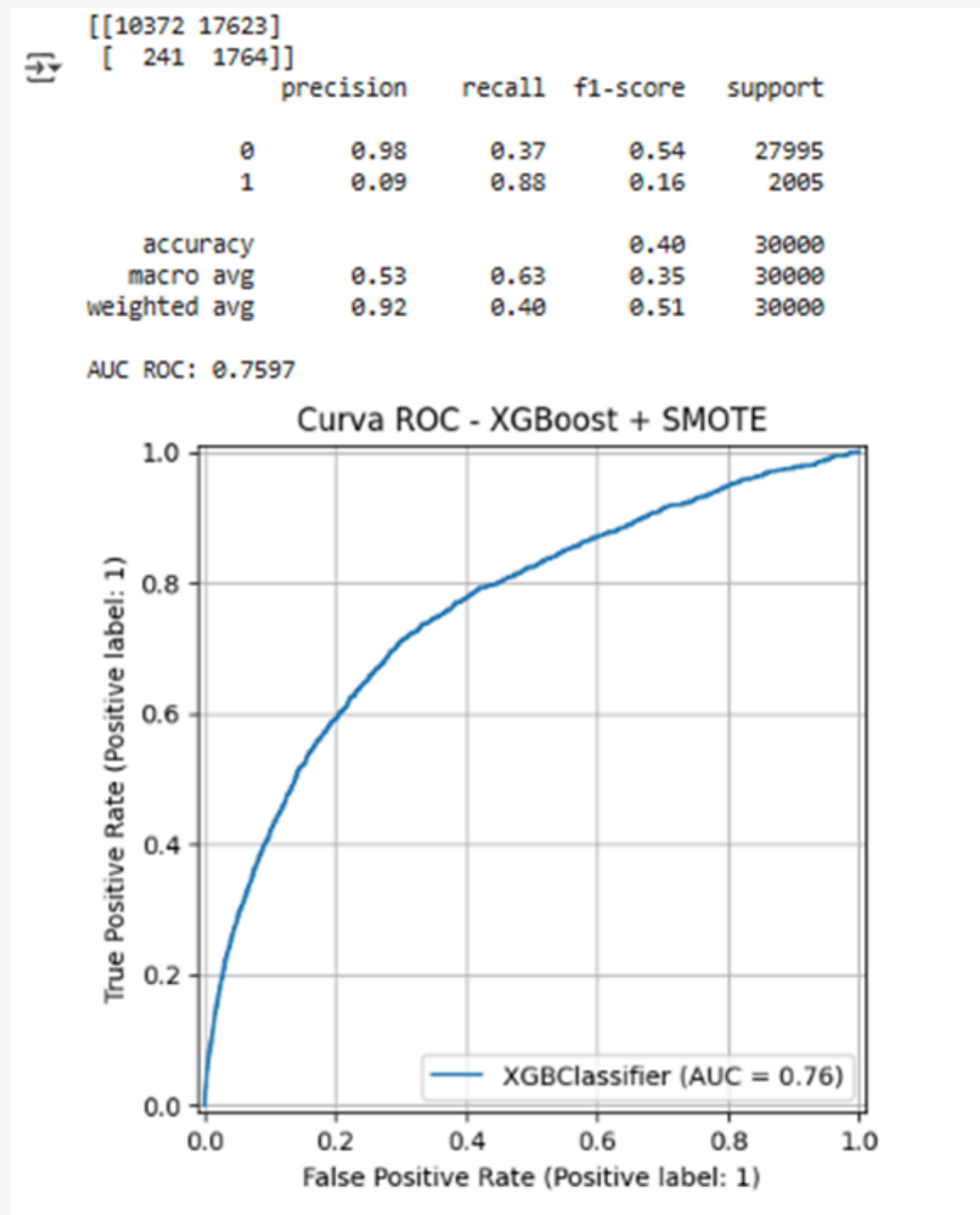
XGBOOST CON SMOTE

Objetivo del modelo:

Corregir el desbalance de clases aplicando la técnica de SMOTE, que genera datos sintéticos de la clase minoritaria (default), y luego aplicar XGBoost, un algoritmo de clasificación avanzado que maneja bien relaciones no lineales y alta dimensionalidad.

MODELOS APLICADOS Y RESULTADOS

XGBOOST CON SMOTE



El modelo detecta correctamente la mayoría de los casos de default (clase 1): 88% de recall.

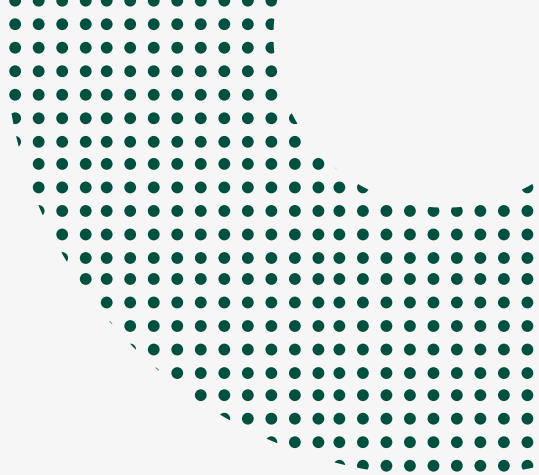
Alta recall en la clase 1 (default) indica que el modelo identifica correctamente a quienes tienen riesgo, aunque a costa de muchos falsos positivos.

El modelo XGBoost con SMOTE logra una importante mejora en la detección de clientes en riesgo de default (recall de 88%), lo que representa un gran avance frente al modelo anterior. Sin embargo, lo hace a costa de una muy baja precisión (9%), es decir, genera muchas falsas alarmas.

A pesar de un accuracy general bajo (40%), el AUC ROC de 0.76 confirma que el modelo tiene un buen poder discriminativo. Esto es coherente con el objetivo de priorizar la detección de riesgo por encima de la exactitud general.

Se puede rechazar parcialmente la hipótesis nula, ya que el modelo sí muestra capacidad predictiva útil sobre la clase de interés, aunque se requiere optimizar el umbral de clasificación o ajustar la precisión mediante otras estrategias.

MODELOS APLICADOS Y RESULTADOS



Variables fundamentales en el proceso

Con base en la importancia de variables en los modelos de árboles y los resultados obtenidos, las variables más relevantes fueron:

Variable	Rol en el modelo	Importancia
Nro prestao retrasados	Historial de mora	Muy alta (más influyente)
Prct uso tc	Uso del crédito disponible	Alta
Prct deuda vs ingresos	Carga financiera total	Alta
Mto ingreso mensual	Capacidad de pago	Media
Nro dependiente	Carga económica del hogar	Media
Edad	Comportamiento financiero según etapa de vida	Menor influencia

CONCLUSIONES:

- **Análisis del experto 1 – Experian (Global Insights Report 2023)**

En su informe global, Experian señala que los factores más determinantes en los modelos de scoring crediticio son el historial de pagos, el uso del crédito disponible (utilización) y el nivel de deuda con respecto a los ingresos. Además, resalta que el uso de modelos basados en aprendizaje automático permite mejorar significativamente la detección temprana del riesgo de incumplimiento.

- Esto coincide con los resultados obtenidos en este proyecto, donde las variables Nro_prestao_retrasados, Prct_uso_tc y Prct_deuda_vs_ingresos resultaron ser las más importantes para predecir el default, y el modelo XGBoost logró una capacidad alta para detectar clientes en riesgo.
- Fuente: Experian Global Insights Report 2023

CONCLUSIONES:

Análisis del experto 2 – Banco Mundial (Financial Inclusion and Credit Access Report)

El Banco Mundial, en sus estudios sobre inclusión financiera, destaca que, aunque factores como el ingreso, la edad y el número de dependientes pueden influir en la capacidad de pago, los factores comportamentales tienen un mayor poder predictivo del riesgo de crédito. En particular, el historial de retrasos y el uso del crédito son los indicadores más confiables para identificar clientes de alto riesgo.

- Los hallazgos del presente análisis respaldan esta perspectiva, ya que las variables demográficas como edad y dependientes tuvieron un impacto menor, mientras que los comportamientos financieros recientes (uso de crédito y mora) fueron claves para detectar incumplimientos.
- Fuente: World Bank - Financial Inclusion & Credit Risk Report