Predicción de Riesgo de Impago Crediticio

Home Credit Default Risk - Un Enfoque Multivariado

Gerardo Guerrero Juan Pablo Cordero Jerónimo Deli Romain S

2025-12-09

Table of contents

1 Introducción

1.1 Contexto del Problema

El acceso al crédito es un pilar fundamental para el desarrollo económico individual y colectivo. Sin embargo, las instituciones financieras enfrentan el desafío constante de evaluar el **riesgo de incumplimiento** (*default*) de sus clientes. Una evaluación inadecuada puede resultar en pérdidas significativas para la institución o, por otro lado, en la exclusión financiera de personas que podrían cumplir con sus obligaciones.

Home Credit Group es una compañía de servicios financieros enfocada en préstamos a poblaciones no bancarizadas o con historial crediticio limitado. El problema que abordamos es la **predicción del riesgo de impago** utilizando técnicas estadísticas multivariadas, con el objetivo de:

1. **Identificar clientes con alta probabilidad de incumplimiento** antes de otorgar el crédito
2. **Comprender los factores que influyen en el impago** para diseñar políticas de mitigación
3. **Equilibrar la inclusión financiera con la gestión del riesgo**

Desde la perspectiva de la gestión de riesgos, el objetivo no es únicamente predecir quién hará o no hará default, sino **estimar probabilidades** que permitan tomar decisiones bajo restricciones de capital, regulación y apetito de riesgo. En este sentido, un modelo de probabilidad de impago se convierte en una herramienta cuantitativa clave para:

- Definir **políticas de originación** (qué tipo de clientes aceptar o rechazar).
- Ajustar **límites de crédito** y condiciones como plazo y tasa.
- Alimentar modelos de **pérdida esperada** (PD $\times$ LGD $\times$ EAD) y de provisiones regulatorias.
- Diseñar estrategias de **seguimiento temprano** (early warning) y cobranza preventiva.

A lo largo del reporte, nuestro énfasis estará en balancear el desempeño estadístico del modelo con su **utilidad práctica** en la toma de decisiones crediticias.

1.2 Marco Conceptual: El Grafo Causal del Impago

Antes de desarrollar nuestros modelos predictivos, construimos un **grafo causal** que representa nuestra comprensión teórica del fenómeno. Este ejercicio de pensamiento causal nos permite identificar las variables relevantes y sus relaciones, fundamentando así nuestro enfoque analítico.

1.2.1 Modelo Causal Simplificado

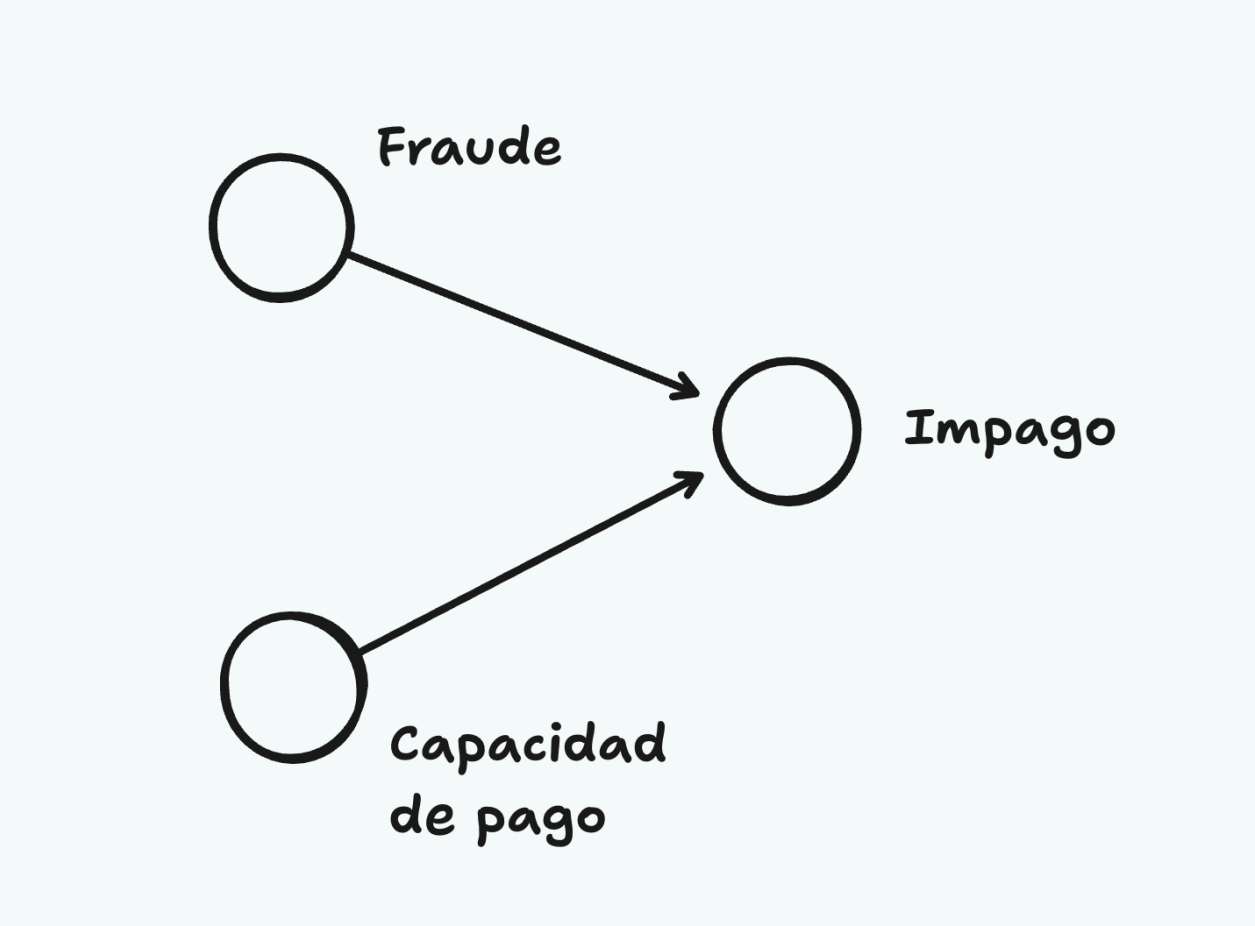


Figure 1: Grafo Causal Simple

El impago crediticio puede originarse por dos vías principales:

- **Fraude**: Cuando el cliente nunca tuvo intención de pagar
- **Capacidad de Pago**: Cuando el cliente no puede cumplir con sus obligaciones debido a restricciones económicas

1.2.2 Modelo Causal Detallado

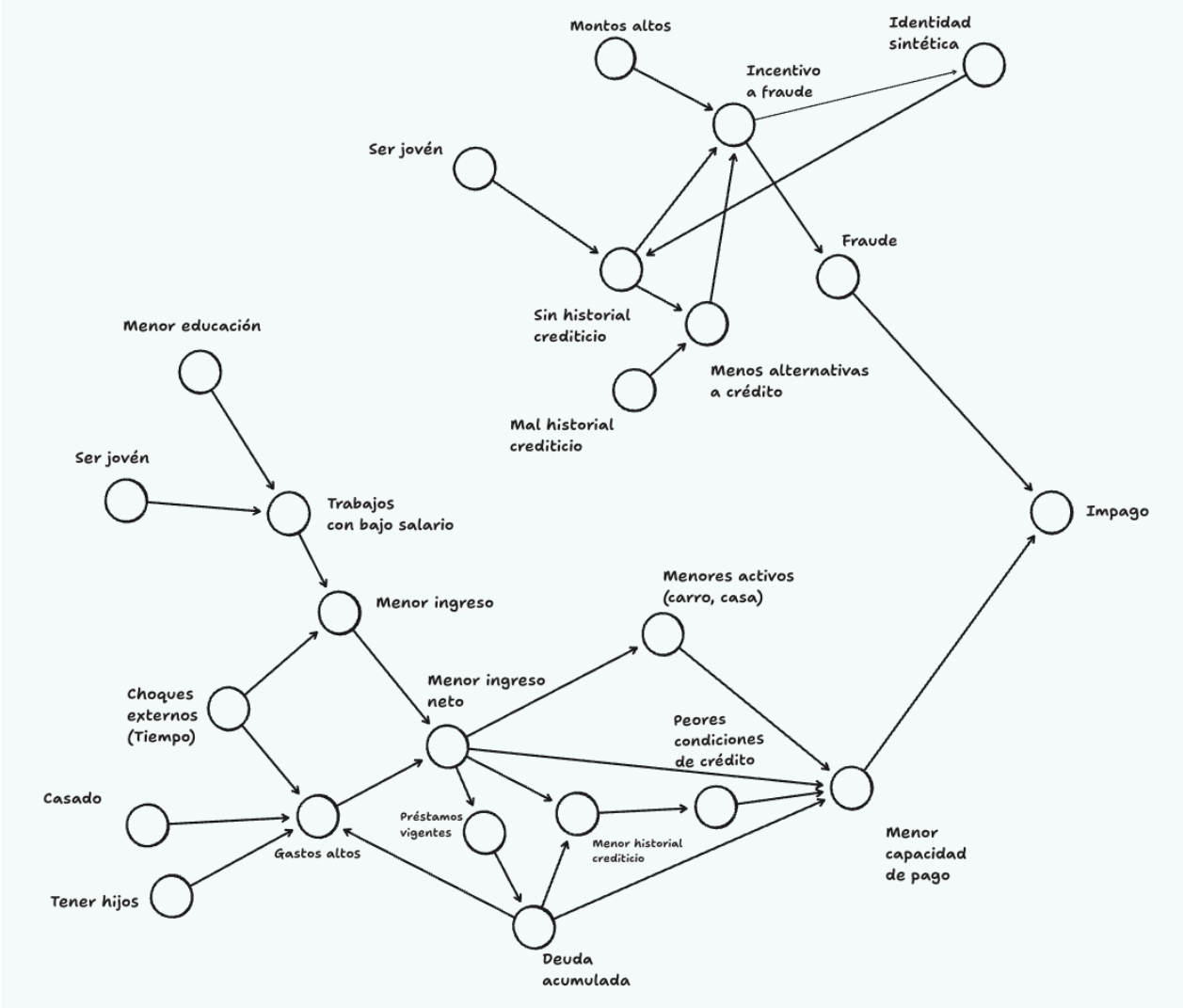


Figure 2: Grafo Causal Detallado

El grafo causal detallado nos muestra las relaciones entre las distintas variables que capturamos en los datos y cómo estas se relacionan con los dos mecanismos principales de impago.

El grafo causal detallado sirve como una **capa intermedia** entre el conocimiento de negocio y la modelación estadística. Más allá de listar variables, nos obliga a responder preguntas como:

- ¿Qué mecanismos generan realmente el impago (fraude vs. incapacidad de pago)?
- ¿Qué variables son las que se correlacionan con impago?
- ¿Qué información proviene del cliente, del buró u otras fuentes externas?

En la práctica, este grafo:

1. **Guía la ingeniería de variables**: por ejemplo, agrupar información de buró en indicadores de mora histórica, carga de deuda y número de créditos activos.
2. **Ayuda a interpretar el modelo a posteriori**: si una variable aparece como importante en el modelo, podemos ubicarla en el grafo y entender si actúa como proxy de fraude, de capacidad de pago o de algún canal más específico.

1.3 Hipótesis de Investigación

Con base en el marco causal, formulamos las siguientes hipótesis que guiarán nuestro análisis:

Hipótesis	Variable Proxy	Relación Esperada
Préstamos más altos incrementan la probabilidad de impago	`AMT_CREDIT`	Positiva
Menor edad y sin historial crediticio aumenta el riesgo	`EDAD_ANOS`, `ES_PRIMER_CREDITO`	Negativa, Positiva
Mal historial crediticio incrementa el riesgo	`EXT_SOURCE_1/2/3`, `SCORE_PROMEDIO`	Negativa
Menor ingreso incrementa el riesgo	`AMT_INCOME_TOTAL`, `INGRESO_PER_CAPITA`	Negativa
Mayor carga de gastos incrementa el riesgo	`CNT_CHILDREN`, `CNT_FAM_MEMBERS`	Positiva
Mayor deuda acumulada incrementa el riesgo	`TOTAL_DEUDA_ACTUAL`, `CREDITOS_ACTIVOS`	Positiva
Menos activos incrementan el riesgo	`NUM_ACTIVOS`, `FLAG_OWN_CAR`, `FLAG_OWN_REALTY`	Negativa
Condiciones crediticias adversas aumentan el riesgo	`TASA_INTERES_PROMEDIO`, `PLAZO_PROMEDIO`	Positiva

1.4 Preguntas de Investigación

1. ¿Cuáles son las variables con mayor poder predictivo para identificar clientes en riesgo de impago?
2. ¿Qué modelo (Regresión Logística, Random Forest o XGBoost) ofrece el mejor balance entre interpretabilidad y poder predictivo?

2 Metodología y Datos

2.1 Descripción del Dataset

El conjunto de datos proviene de la competencia Home Credit Default Risk de Kaggle. La estructura de datos incluye múltiples tablas relacionadas:

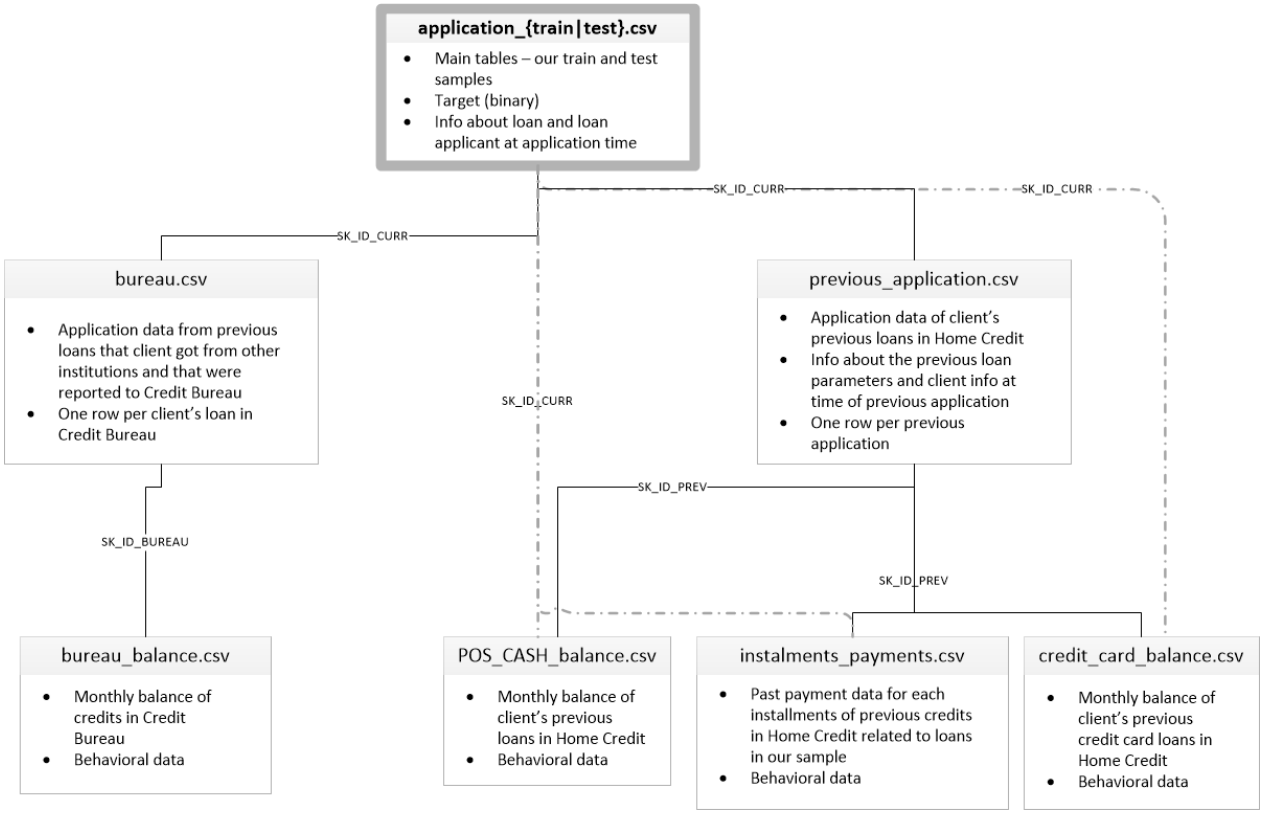


Figure 3: dataset-description

Una característica importante de este dataset es su **estructura relacional**. La tabla `application_train` contiene la observación "principal" (cada fila es una solicitud de crédito), mientras que el resto de tablas describen el **historial del cliente en distintas dimensiones**:

- `bureau` y `bureau_balance` capturan el comportamiento del cliente en otras instituciones financieras.
- `previous_application`, `installments_payments` y `credit_card_balance` describen el ciclo de vida de créditos previos con Home Credit (originación, pagos, atrasos, etc.).
- `POS_CASH_balance` complementa la información con productos tipo POS / cash loans.

Esto implica dos retos principales:

1. **Integración de información**: es necesario definir una estrategia para pasar de tablas transaccionales (múltiples registros por cliente) a variables agregadas por `SK_ID_CURR`.
2. **Escalabilidad computacional**: algunas tablas tienen millones de filas; decisiones de agregación, muestreo y tipos de datos afectan directamente el tiempo de cómputo y el uso de memoria.

La ingeniería de variables que describimos en la siguiente sección responde precisamente a estos retos.

2.2 Carga de Datos

Table 2: Estructura del dataset Home Credit Default Risk

Table 2

Tabla	Descripción	Registros
application_train.csv	Información principal de las solicitudes	307,511
bureau.csv	Historial crediticio de otras instituciones	1,716,428

Tabla	Descripción	Registros
bureau_balance.csv	Balance mensual de créditos en buró	27,299,925
previous_application.csv	Solicitudes previas en Home Credit	1,670,214
installments_payments.csv	Historial de pagos	13,605,401
credit_card_balance.csv	Balance de tarjetas de crédito	3,840,312

2.3 Variables Originales de Application Train

La tabla principal `application_train.csv` contiene 122 variables que se pueden agrupar en las siguientes categorías (mostramos las más relevantes):

2.3.1 Variables Demográficas

- `CODE_GENDER`: Género del solicitante
- `DAYS_BIRTH`: Edad en días (negativo)
- `NAME_FAMILY_STATUS`: Estado civil
- `CNT_CHILDREN`: Número de hijos
- `CNT_FAM_MEMBERS`: Número de miembros en la familia
- `NAME_EDUCATION_TYPE`: Nivel educativo
- `NAME_INCOME_TYPE`: Tipo de ingreso

2.3.2 Variables Financieras

- `AMT_INCOME_TOTAL`: Ingreso total del solicitante
- `AMT_CREDIT`: Monto del crédito solicitado
- `AMT_ANNUITY`: Anualidad del préstamo
- `AMT_GOODS_PRICE`: Precio del bien a comprar

2.3.3 Scores de Riesgo Externos

- `EXT_SOURCE_1`: Score buro de crédito 1
- `EXT_SOURCE_2`: Score buro de crédito 2
- `EXT_SOURCE_3`: Score buro de crédito 3

2.3.4 Variables de Activos

- `FLAG_OWN_CAR`: Si posee automóvil
- `FLAG_OWN_REALTY`: Si posee propiedad inmobiliaria
- `OWN_CAR_AGE`: Edad del automóvil

3 Ingeniería de Variables

3.1 Generación de Features

A partir de las tablas relacionadas, construimos variables que capturan diferentes dimensiones del riesgo crediticio:

La lógica de la ingeniería de variables que implementamos se puede resumir en cuatro bloques principales:

1. **Perfil socio-demográfico y capacidad de pago**
   A partir de `application_train` construimos indicadores de edad (`EDAD_ANOS`), estructura familiar (`CNT_CHILDREN`, `CNT_FAM_MEMBERS`) e ingresos (`AMT_INCOME_TOTAL`, `INGRESO_PER_CAPITA`). Estos proxies buscan capturar la **capacidad de generar flujo de efectivo** para servir la deuda.

2. **Carga de deuda y uso del crédito**
   De las tablas de buró (`bureau`, `bureau_balance`) extraemos el nivel de crédito actual (`TOTAL_DEUDA_ACTUAL`, `TOTAL_CREDITO_OTORGADO`), el número de créditos activos (`CREDITOS_ACTIVOS`) y la historia de mora (`PCT_MESES_MORA`, `CREDITOS_CON_IMPAGO`). Estas variables resumen el **apetito de endeudamiento** y el cumplimiento pasado del cliente.

3. **Condiciones crediticias históricas**
   De `previous_application` construimos variables como `TASA_INTERES_PROMEDIO`, `PLAZO_PROMEDIO` y `TOTAL_CREDITO_HISTORICO`. El objetivo es capturar el tipo de condiciones bajo las cuales el cliente ha tomado crédito en el pasado, lo cual puede estar correlacionado con su riesgo (p.ej., tasas más altas pueden ser reflejo de mayor riesgo percibido).

4. **Comportamiento de pago reciente**
   A partir de `installments_payments` y `credit_card_balance` construimos ratios como `RATIO_PAGO_CUOTA` y `RATIO_PAGO_MINIMO_TC`, que miden qué tan sistemáticamente el cliente paga sus cuotas completas o sólo mínimos. Estas son variables de **conducta reciente**, muy relevantes para anticipar problemas de liquidez.

En conjunto, el dataset final combina información estática (perfil del cliente al momento de la solicitud) con información histórica dinámica (trayectoria de uso y pago de crédito), alineado con el grafo causal propuesto.

Una observación importante es que muchas de las variables nuevas presentan valores faltantes (`NaN`). Esto puede deberse a que ciertos clientes no tienen historial en buró o no han tenido créditos previos, lo que limita la información disponible para su análisis.

3.2 Distribución de la Variable Objetivo

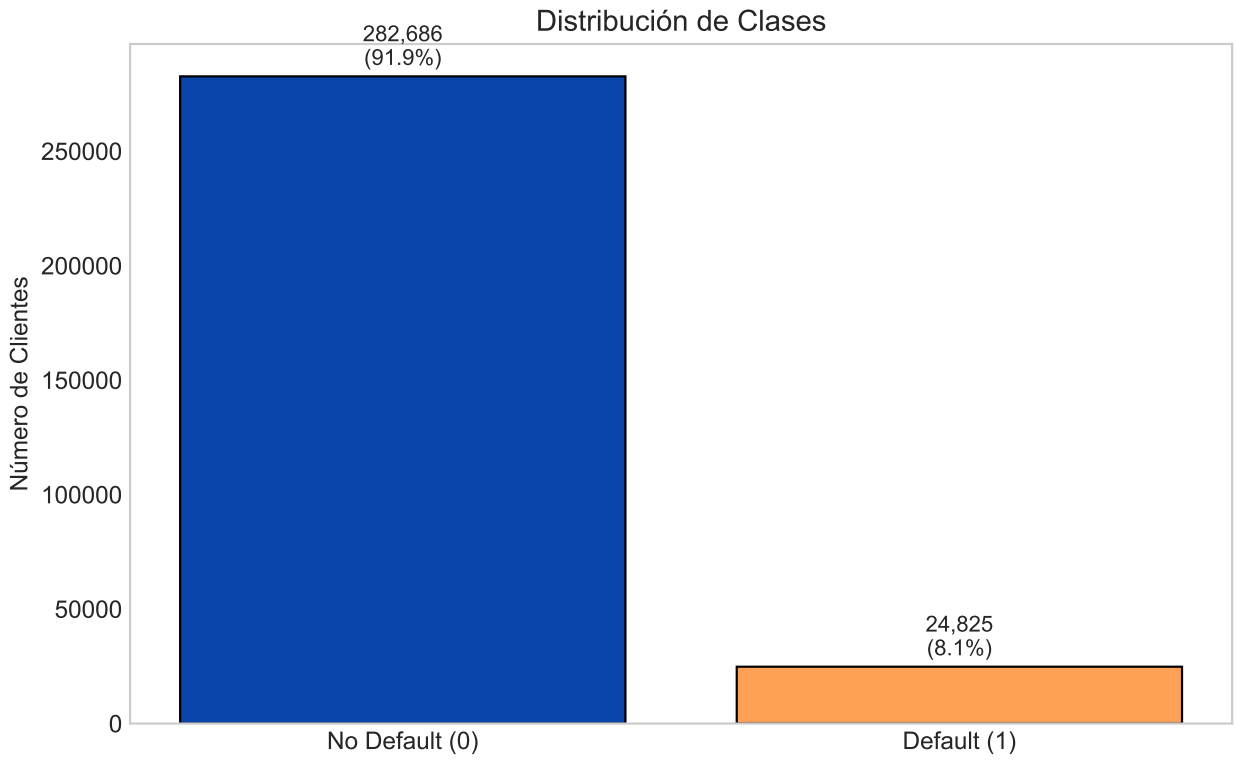


Figure 4: Distribución de la variable TARGET (desbalance de clases)

La tasa de default observada (~8%) es consistente con portafolios de consumo relativamente amplios y diversificados. Desde el punto de vista de modelación, este **desbalance de clases** implica que:

- Un modelo que ignore la clase minoritaria puede mostrar métricas aparentemente buenas (p. ej. alta accuracy), pero ser inútil para **detectar clientes realmente riesgosos**.
- La función de pérdida "natural" del modelo está desbalanceada: equivocarse en un cliente "default" (falso negativo) es mucho más costoso que equivocarse en un "no default".

Por ello, adoptamos dos estrategias complementarias:

1. Ajustar el modelo con pesos de clase (`class_weight='balanced'` o `scale_pos_weight`), de modo que los errores en la clase positiva tengan mayor penalización.
2. Enfocarnos en métricas más sensibles al desbalance, como **PR-AUC** (Average Precision), **Recall** y análisis por **deciles de score**, en lugar de depender únicamente de ROC-AUC o accuracy.

3.3 Análisis de Correlaciones

El análisis de correlaciones muestra que muchas variables están fuertemente relacionadas entre sí, especialmente:

- Montos monetarios derivados de la misma base (`AMT_CREDIT`, `AMT_ANNUITY`, `AMT_GOODS_PRICE`, etc.).
- Scores externos (`EXT_SOURCE_1/2/3`) que, aunque provienen de fuentes distintas, tienden a moverse juntos.

En modelos lineales como la regresión logística, la **multicolinealidad** puede inflar varianzas de los coeficientes y dificultar la interpretación. Aunque los modelos de árboles son más robustos a este problema, mantener demasiadas variables redundantes puede:

- Introducir ruido innecesario.
- Aumentar el costo computacional.
- Complicar la comunicación de resultados.

Por ello, en la etapa de **selección de variables** optamos por:

- Conservar métricas sintéticas más interpretables (p.ej. `SCORE_PROMEDIO`, `CREDIT_INCOME_RATIO`).
- Eliminar componentes redundantes (p.ej. `EXT_SOURCE_1/2/3` por separado, o montos crudos que ya están en ratios).
- Mantener un subconjunto manejable de variables con significado económico claro, manteniendo el foco en **calidad sobre cantidad** de features.
- Hicimos un análisis de VIF (Variance Inflation Factor) de forma iterativa para eliminar variables con alta multicolinealidad.

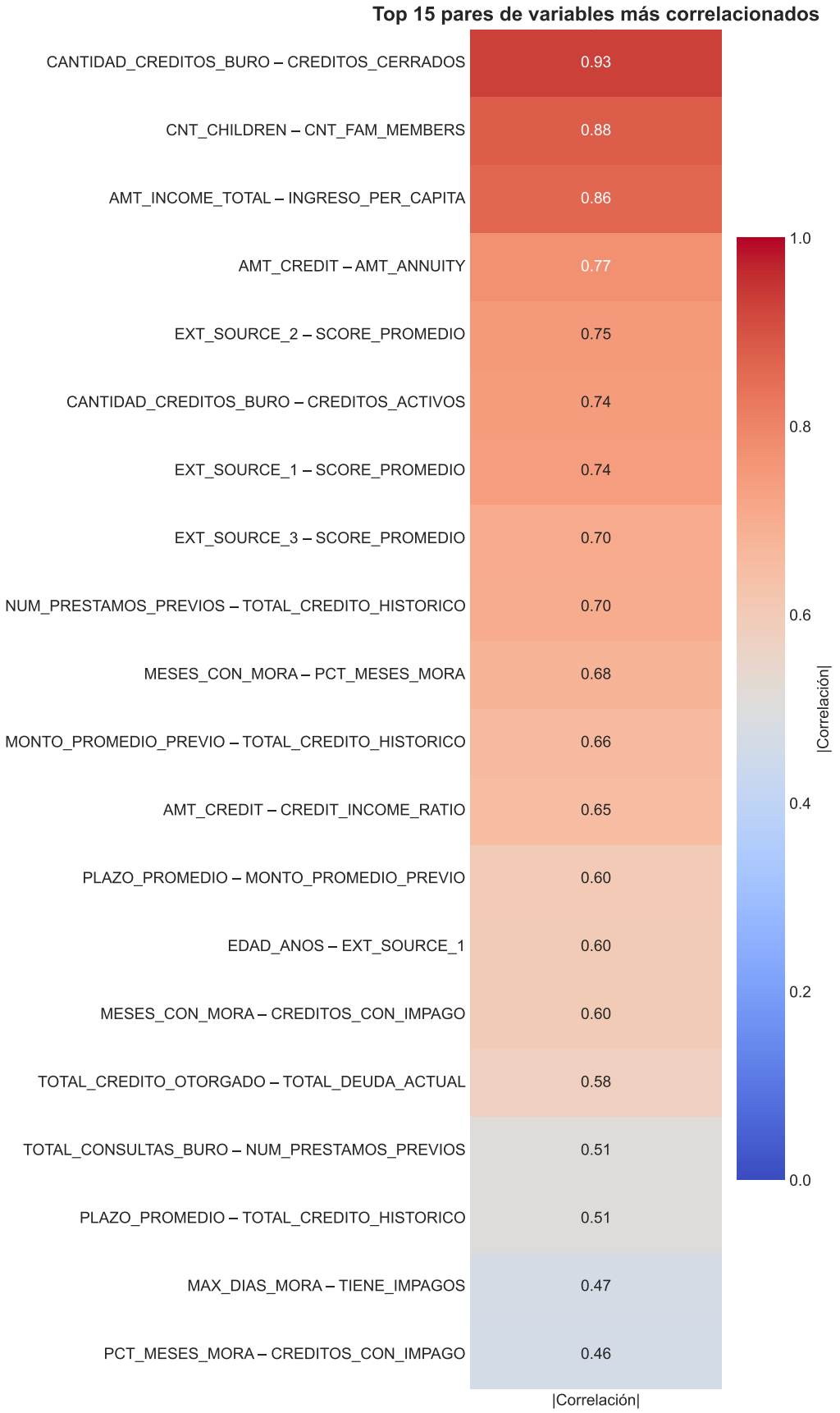


Figure 5: Heatmap de correlaciones entre variables numéricas

3.4 Selección de Variables

Eliminamos variables redundantes para reducir la multicolinealidad:

```
Variables restantes: 25
Observaciones: 307,511

Variables en el modelo final:
  1. EDAD_ANOS
  2. SCORE_PROMEDIO
  3. CREDIT_INCOME_RATIO
  4. NAME_FAMILY_STATUS
  5. CNT_CHILDREN
  6. CODE_GENDER
  7. NAME_EDUCATION_TYPE
  8. INGRESO_PER_CAPITA
  9. NUM_ACTIVOS
  10. TOTAL_CONSULTAS_BURO
  11. TOTAL_CREDITO_DISPONIBLE
  12. TOTAL_CREDITO_OTORGADO
  13. TOTAL_DEUDA_ACTUAL
  14. MAX_DIAS_MORA
  15. CREDITOS_ACTIVOS
  16. CREDITOS_CERRADOS
  17. PCT_MESES_MORA
  18. CREDITOS_CON_IMPAGO
  19. NUM_PRESTAMOS_PREVIOS
  20. TASA_INTERES_PROMEDIO
  21. PLAZO_PROMEDIO
  22. MONTO_PROMEDIO_PREVIO
  23. TOTAL_CREDITO_HISTORICO
  24. RATIO_PAGO_CUOTA
  25. RATIO_PAGO_MINIMO_TC
```

4 Regresión Logística

4.1 Teoría

4.1.1 Función Sigmoide

La regresión logística es un modelo de clasificación que estima la probabilidad de que una observación pertenezca a una clase particular. El modelo utiliza la **función sigmoide** (o logística) para transformar una combinación lineal de las variables predictoras en una probabilidad:

$$P(Y = 1|X) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Donde:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

4.1.2 Interpretación de Coeficientes

Los coeficientes β_j se interpretan en términos de **odds ratios**:

$$\text{OR}_j = e^{\beta_j}$$

- Si $\beta_j > 0$: La variable aumenta la probabilidad de default
- Si $\beta_j < 0$: La variable disminuye la probabilidad de default
- Si $\beta_j = 0$: La variable no tiene efecto

4.1.3 Función de Pérdida: Log-Loss (Entropía Cruzada Binaria)

Los parámetros se estiman minimizando la **log-loss** (también conocida como entropía cruzada binaria):

$$\mathcal{L}(\beta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{p}_i) + (1-y_i)\log(1-\hat{p}_i)]$$

4.1.4 Regularización

Para prevenir el overfitting, aplicamos **regularización L2 (Ridge)**:

$$\mathcal{L}_{reg}(\beta) = \mathcal{L}(\beta) + \lambda\sum_{j=1}^{p}\beta_j^2$$

4.2 Entrenamiento del Modelo

```
================================================================================
MODELO DE REGRESION LOGISTICA - HOME CREDIT DEFAULT RISK
================================================================================

Preparando datos...
Variables predictoras: 25
Variables numericas: 22
Variables categoricas: 3

Imputando valores faltantes...
Codificando variables categoricas...

Dividiendo datos en Train/Test...
Train set: 246,008 usuarios (80.0%)
Test set:  61,503 usuarios (20.0%)

Normalizando variables (StandardScaler)...

Entrenando Regresion Logistica...
  Modelo entrenado exitosamente

Validacion Cruzada (5-fold)...
ROC-AUC por fold: [0.73161298 0.72606379 0.73067105 0.73487217 0.73885166]
Media: 0.7324 (+/- 0.0043)
```

4.3 Resultados del Modelo

Table 3: Métricas del modelo de Regresión Logística

Table 3

Métrica	Test
ROC-AUC	0.7337

Métrica	Test
Average Precision	0.2109
Accuracy	0.6841

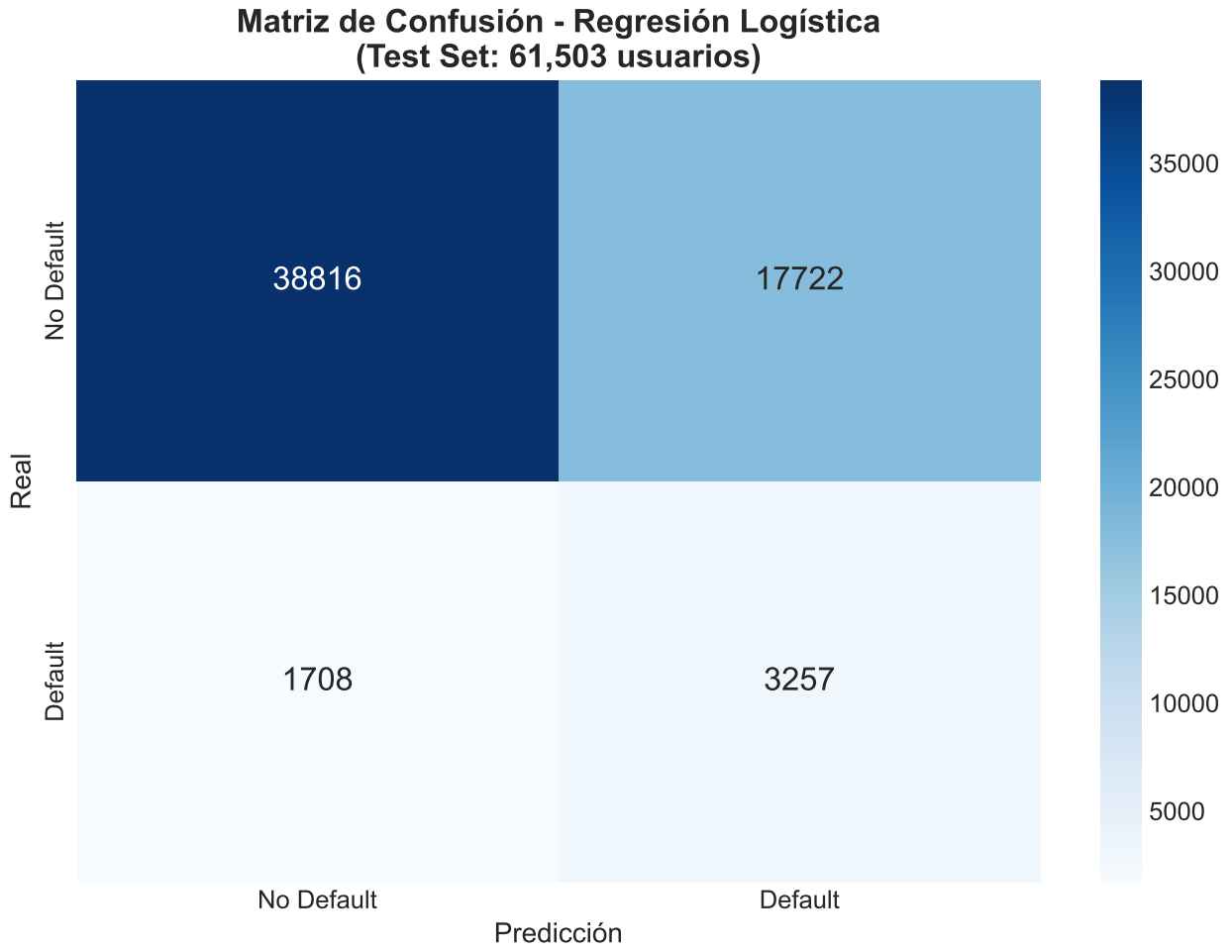


Figure 6: Matriz de Confusión - Regresión Logística (Test Set)

```
Interpretación:
  Verdaderos Negativos (TN): 38,816 - Predijo 'Pago' y si pagó
  Falsos Positivos (FP): 17,722 - Predijo 'Default' pero pagó
  Falsos Negativos (FN): 1,708 - Predijo 'Pago' pero hizo default
  Verdaderos Positivos (TP): 3,257 - Predijo 'Default' y si hizo default
```

4.3.1 Curvas ROC y Precision-Recall

En el caso de la regresión logística, el ROC-AUC y el Average Precision en el set de prueba indican que el modelo logra **discriminar razonablemente bien** entre clientes que harán default y aquellos que no, aun en presencia de fuerte desbalance. El hecho de que las métricas de entrenamiento y prueba sean cercanas sugiere que el modelo **no presenta overfitting severo**, lo cual es coherente con la regularización L2 y la relativa parsimonia del conjunto de variables.

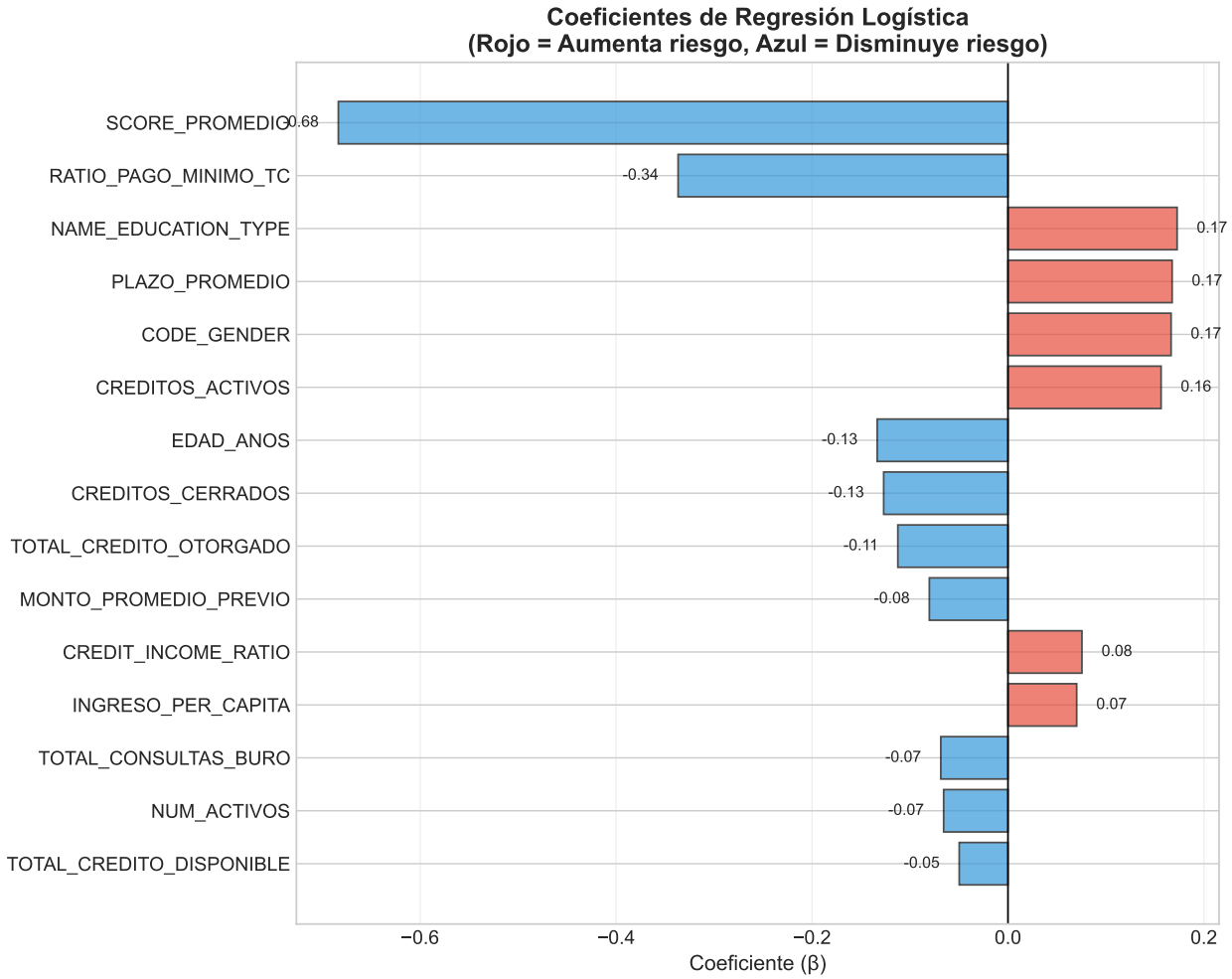


Figure 7: Coeficientes de Regresión Logística (Top 15 variables)

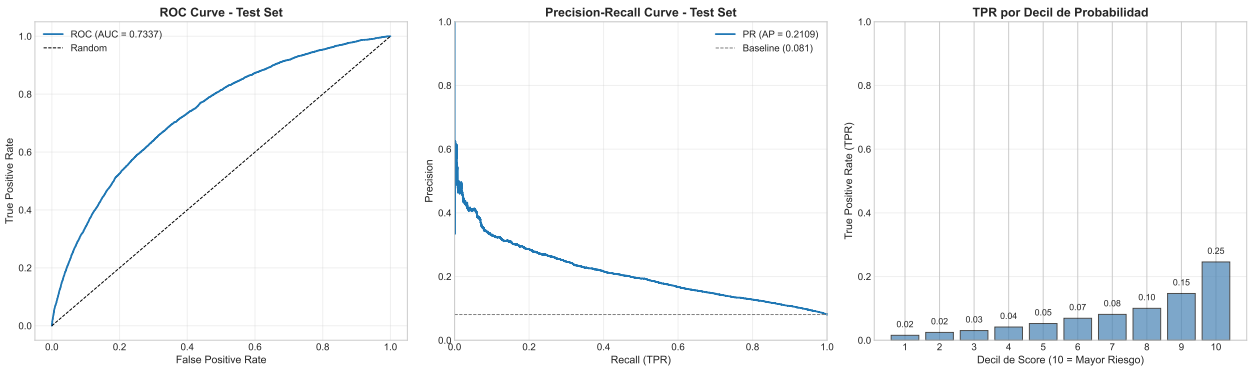


Figure 8: Curvas ROC y Precision-Recall - Regresión Logística

Dado que se trata de un modelo lineal, la información contenida en los coeficientes se puede traducir a **insights de negocio**:

- Coeficientes positivos indican variables que **incrementan** el riesgo de impago (por ejemplo, mayores ratios de deuda o mayor proporción de meses con mora).
- Coeficientes negativos corresponden a variables que **reducen** el riesgo (p. ej. mayores ingresos per cápita o mejores scores externos).

Este modelo, aun cuando no es el de mejor desempeño, proporciona una **línea base interpretable** contra la cual podemos comparar modelos más complejos.

La matriz de confusión permite visualizar con claridad los **errores de clasificación**:

- Los **falsos negativos (FN)** corresponden a clientes que el modelo considera "buenos" pero que terminan en default; son los más costosos desde el punto de vista del riesgo.
- Los **falsos positivos (FP)** son clientes rechazados o tratados como de alto riesgo, aunque finalmente hubieran pagado; reflejan un costo de **oportunidad** y posible pérdida de negocio.

5 Random Forest

5.1 Descripción del Modelo

Random Forest es un algoritmo de **ensemble learning** basado en árboles de decisión. El modelo construye múltiples árboles utilizando:

1. **Bagging (Bootstrap Aggregating)**: Cada árbol se entrena con una muestra bootstrap del conjunto de datos
2. **Selección aleatoria de features**: En cada split, solo se considera un subconjunto aleatorio de variables

Para clasificación, la predicción final se obtiene por **votación mayoritaria**:

$$\hat{y} = \text{mode}\{\hat{y}_1, \hat{y}_2, ..., \hat{y}_B\}$$

Donde B es el número de árboles.

5.1.1 Importancia de Variables (Gini Importance)

La importancia de una variable se mide como la reducción promedio de la impureza de Gini:

$$\text{Gini}(t) = 1 - \sum_{k=1}^{K} p_{tk}^2$$

Donde p_{tk} es la proporción de observaciones de clase k en el nodo t.

5.2 Entrenamiento del Modelo

```
================================================================================
MODELO DE RANDOM FOREST - HOME CREDIT DEFAULT RISK
================================================================================

Train set: 246,008 usuarios
Test set:  61,503 usuarios

Entrenando Random Forest...
  Modelo entrenado exitosamente
```

```
  Número de árboles: 100
  Profundidad máxima: 15

Validacion Cruzada (5-fold)...
ROC-AUC por fold: [0.7399194  0.72517727 0.73832354 0.73798331 0.74308408]
Media: 0.7369 (+/- 0.0061)
```

5.3 Resultados del Modelo

Table 4: Métricas del modelo Random Forest

Table 4

Métrica	Test
ROC-AUC	0.739559
Average Precision	0.215535
Accuracy	0.816968
Overfitting (Gap)	0.172557

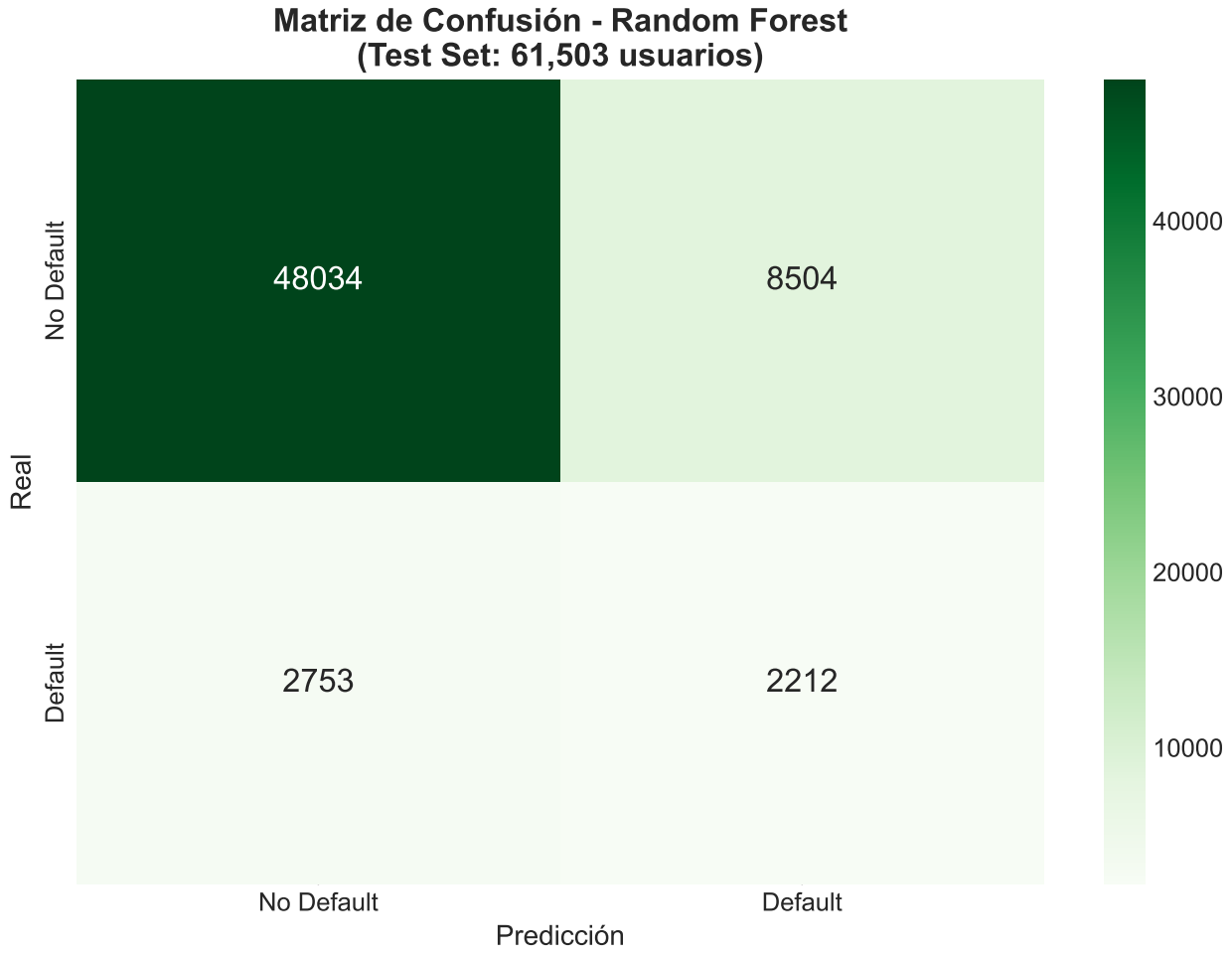
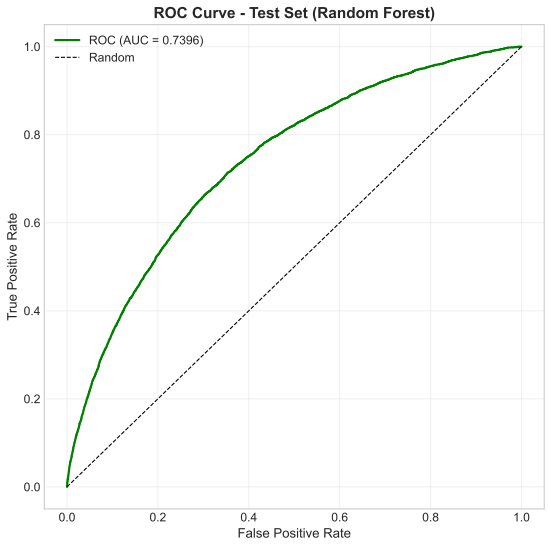
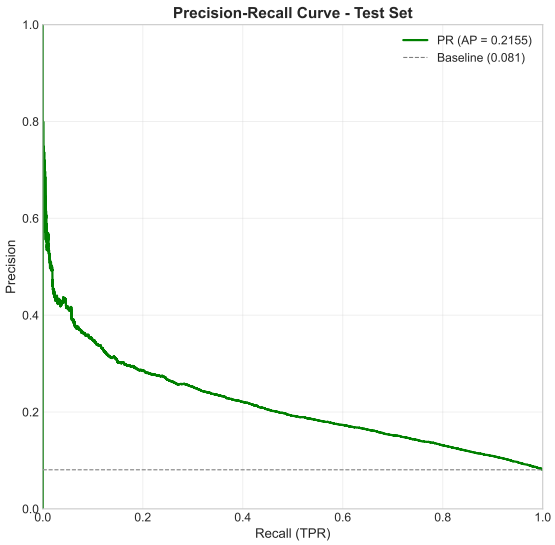
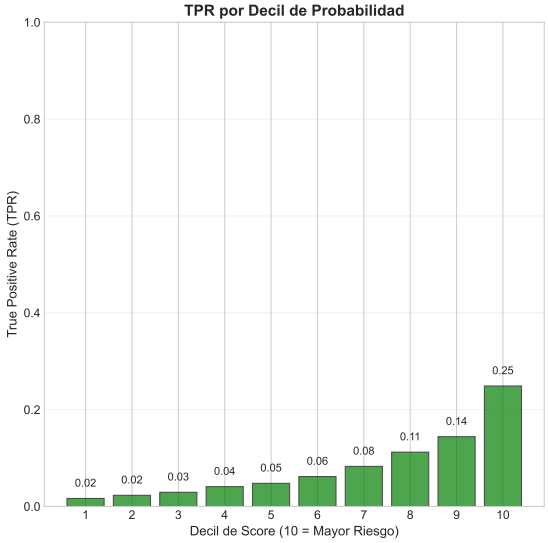
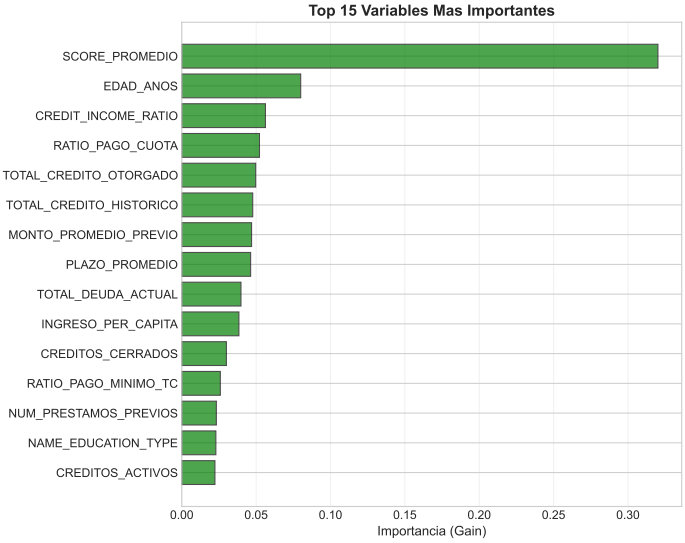


Figure 9: Matriz de Confusión - Random Forest (Test Set)









6 XGBoost

6.1 Descripción del Modelo

XGBoost (Extreme Gradient Boosting) es un algoritmo de **boosting** que construye árboles de decisión de manera secuencial. A diferencia de Random Forest que construye árboles en paralelo, XGBoost:

1. **Entrena árboles secuencialmente**: Cada árbol corrige los errores del anterior
2. **Utiliza gradient boosting**: Optimiza una función de pérdida usando gradiente descendente
3. **Incluye regularización**: Penaliza la complejidad del modelo para prevenir overfitting

6.1.1 Mecanismo de Boosting

El modelo final es una suma ponderada de árboles:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

Donde cada árbol f_k se entrena para minimizar:

$$\mathcal{L}^{(k)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(k-1)} + f_k(x_i)) + \Omega(f_k)$$

La regularización $\Omega(f_k)$ incluye: - **L1 (alpha)**: Regularización de pesos - **L2 (lambda)**: Regularización de scores de hojas - **gamma**: Penalización por complejidad del árbol

6.1.2 Learning Rate ( )

El learning rate controla la contribución de cada árbol:

$$\hat{y}_i^{(k)} = \hat{y}_i^{(k-1)} + \eta \cdot f_k(x_i)$$

6.2 Entrenamiento del Modelo

```
================================================================================
MODELO XGBOOST - HOME CREDIT DEFAULT RISK
================================================================================

Train set: 246,008 usuarios
Test set:  61,503 usuarios

Scale_pos_weight (ratio negativo/positivo): 11.39

Entrenando XGBoost...
  Modelo entrenado exitosamente
  Número de árboles: 100
  Profundidad máxima: 6
  Learning rate: 0.1

Validacion Cruzada (5-fold)...
ROC-AUC por fold: [0.75090577 0.73973452 0.74770615 0.75076835 0.75653203]
Media: 0.7491 (+/- 0.0055)
```

6.3 Resultados del Modelo

Table 5: Métricas del modelo XGBoost

Table 5

Métrica	Test
ROC-AUC	0.752887
Average Precision	0.233792
Accuracy	0.707022
Overfitting (Gap)	0.042607

6.3.1 Matriz de Confusión

6.3.2 Importancia de Variables

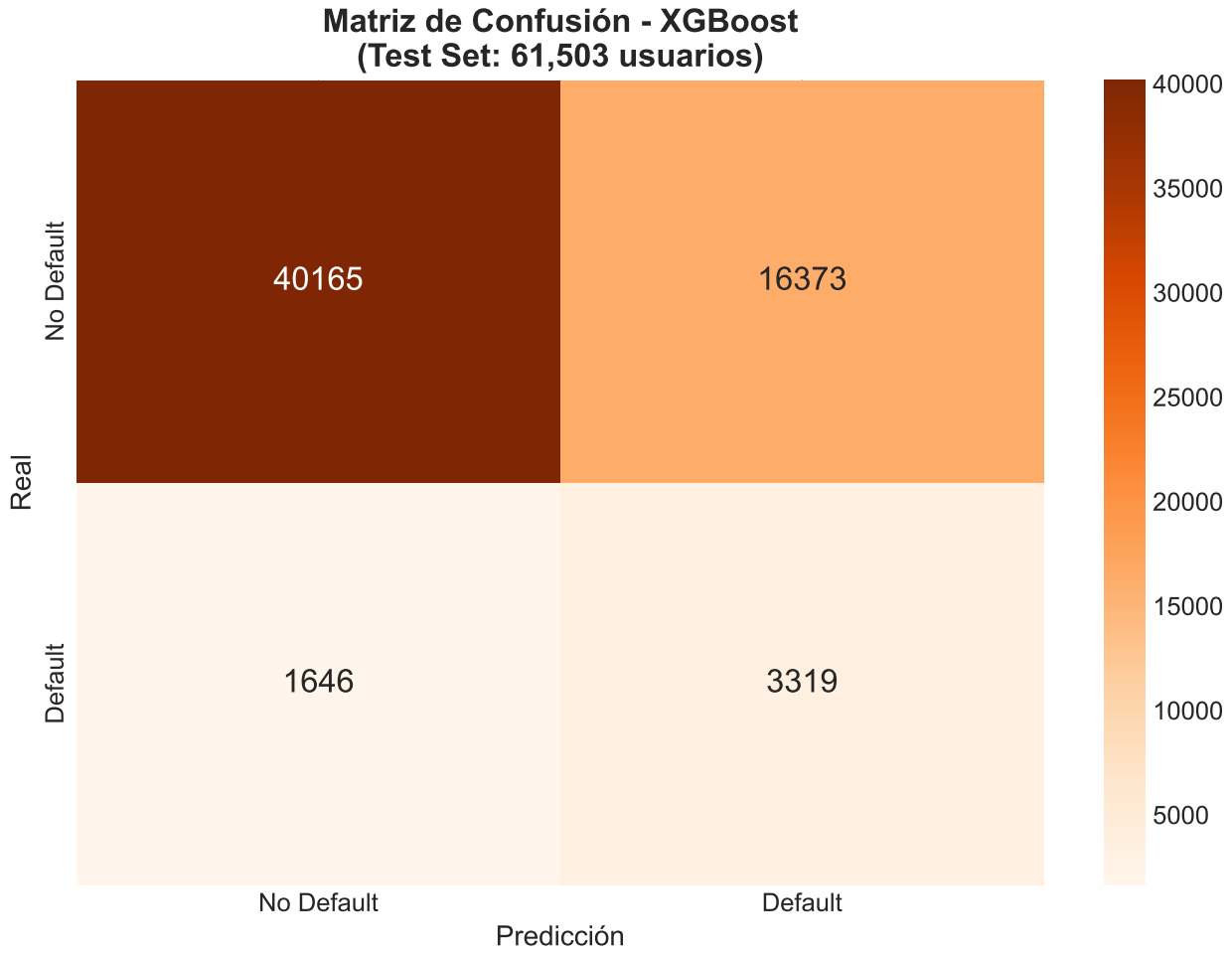
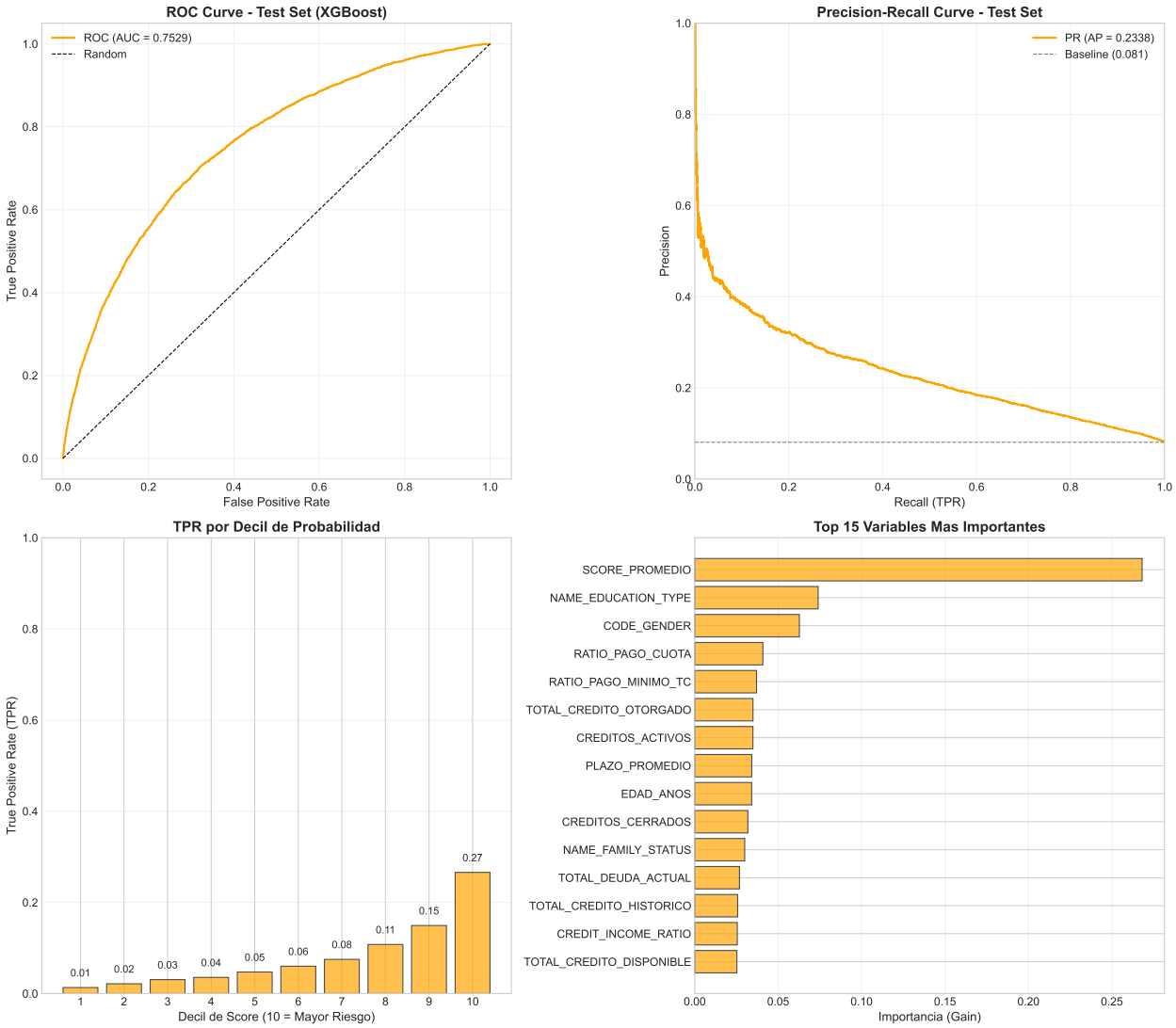


Figure 10: Matriz de Confusión - XGBoost (Test Set)



Tanto Random Forest como XGBoost logran mejorar las métricas de desempeño respecto a la regresión logística, lo cual era esperable dado que:

- Capturan **relaciones no lineales** entre las variables.
- Modelan de manera más flexible interacciones entre features (por ejemplo, combinaciones de score, carga de deuda y edad).

En el caso de Random Forest, la diferencia entre el desempeño en entrenamiento y prueba es mayor, lo que indica cierto **overfitting** inherente al modelo. XGBoost, por su parte, muestra un mejor balance entre ajuste y generalización, gracias a:

- La regularización explícita en la función de pérdida.
- El uso de un learning rate controlado.
- La posibilidad de ajustar hiperparámetros finos (profundidad, min_child_weight, subsample, etc.).

En resumen, mientras que Random Forest es un buen baseline no lineal relativamente fácil de configurar, XGBoost ofrece un mayor **potencial de performance** a cambio de mayor complejidad de tuning. Es interesante notar que, pese a las diferencias de arquitectura, los tres modelos coinciden en resaltar un conjunto relativamente reducido de variables como las más importantes:

- `SCORE_PROMEDIO` y otras variables relacionadas con el **historial crediticio externo**.
- Ratios que combinan deuda e ingreso (`CREDIT_INCOME_RATIO`, `RATIO_PAGO_CUOTA`).

- Indicadores de comportamiento de pago (`PCT_MESES_MORA`, `CREDITOS_CON_IMPAGO`).

Esta coincidencia refuerza la idea de que el riesgo de impago está fuertemente determinado por una combinación de **historial de cumplimiento** y **capacidad de pago actual**, en línea con la narrativa del grafo causal.

7 Comparación de Modelos

7.1 Métricas Globales

Table 6: Resumen comparativo de todos los modelos

Table 6

Métrica	Regresión Logística	Random Forest	XGBoost
ROC-AUC (Test)	0.7337	0.7396	0.7529
Average Precision (Test)	0.2109	0.2155	0.2338
Accuracy (Test)	0.6841	0.8170	0.7070
ROC-AUC (Train)	0.7329	0.9121	0.7955
Overfitting (Gap)	0.0009	0.1726	0.0426

7.2 Curvas ROC Comparativas

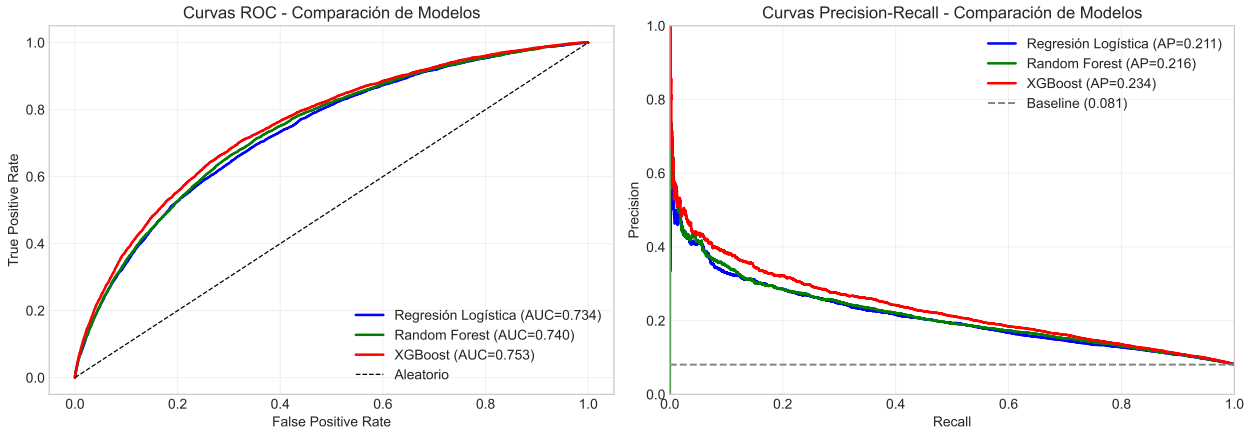


Figure 11: Comparación de curvas ROC entre modelos

7.3 Matrices de Confusión Comparativas

Desde una perspectiva práctica, la elección del modelo no debería basarse únicamente en el valor numérico de AUC o AP, sino también en:

1. **Estabilidad temporal**: qué tan robusto es el modelo frente a cambios en la distribución de clientes y condiciones macroeconómicas.
2. **Implementabilidad**: facilidad de desplegar el modelo en sistemas productivos, tiempos de scoring, dependencia de librerías, etc.
3. **Gobernanza y explicabilidad**: requisitos regulatorios o internos que demanden cierta transparencia en la toma de decisiones.

En ese sentido:

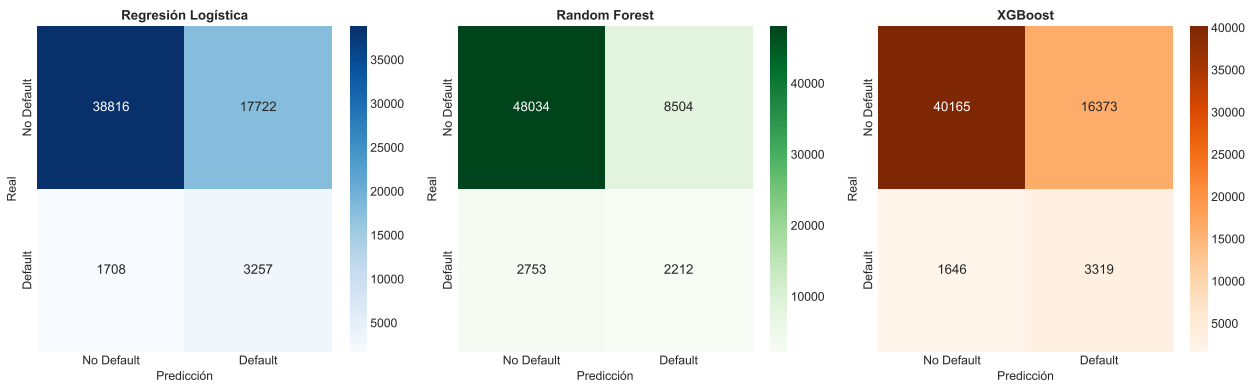


Figure 12: Matrices de confusión de los tres modelos (Test Set)

- La **regresión logística** es atractiva para documentación, auditoría y explicación a no-técnicos.
- **Random Forest** ofrece una mejora de desempeño, pero puede ser más difícil de explicar y controlar si se usan muchos árboles profundos.
- **XGBoost** se perfila como el candidato más competitivo para producción, siempre que se acompañe de herramientas de explicabilidad que permitan entender contribuciones a nivel cliente.

8 Conclusiones

8.1 Hallazgos Principales

8.1.1 1. El Score Crediticio es el Predictor Dominante

En los tres modelos, `SCORE_PROMEDIO` emerge como la variable más importante. Esto confirma que los scores de fuentes externas capturan información valiosa sobre el riesgo crediticio.

8.1.2 2. Comparación de Rendimiento

`Mejor modelo por ROC-AUC: XGBoost (0.7529)`

`Resumen:`

- `Regresión Logística: ROC-AUC = 0.7337, Overfitting = 0.0009`
- `Random Forest: ROC-AUC = 0.7396, Overfitting = 0.1726`
- `XGBoost: ROC-AUC = 0.7529, Overfitting = 0.0426`

8.1.3 3. Trade-off entre Interpretabilidad y Rendimiento

- **Regresión Logística**: Mayor interpretabilidad (coeficientes directos), menor rendimiento
- **XGBoost**: Mejor rendimiento, pero menor interpretabilidad
- **Random Forest**: Balance intermedio, pero con mayor overfitting

8.2 Recomendación Final

8.2.1 Para Producción

Recomendamos implementar **XGBoost** como modelo principal por:

1. **Mejor rendimiento predictivo**
2. **Buen balance recall-precision**
3. **Bajo overfitting**
4. **Robustez ante desbalance de clases**

8.2.2 Validación de Hipótesis

Hipótesis	Resultado	Evidencia
Menor score crediticio → Mayor riesgo	Confirmada	Variable #1 en todos los modelos
Menor edad → Mayor riesgo	Confirmada	Coeficiente negativo en LR
Mayor historial de mora → Mayor riesgo	Confirmada	PCT_MESES_MORA significativo
Más créditos activos → Mayor riesgo	Confirmada	Coeficiente positivo en LR

8.3 Limitaciones y Trabajo Futuro

Aunque los resultados son prometedores, nuestro análisis presenta algunas limitaciones que abren la puerta a extensiones interesantes:

- **Costos y beneficios explícitos**: no incorporamos un análisis formal de costos de FN/FP ni un modelo de utilidad esperada. Incluir una matriz de costos permitiría **optimizar directamente decisiones de originación** en lugar de solo métricas estadísticas.
- **Drift y monitoreo**: en operación real, es fundamental establecer un esquema de monitoreo de **drift de datos y performance** para detectar deterioros en la capacidad predictiva del modelo.

A pesar de estas limitaciones, el ejercicio demuestra cómo, a partir de un marco causal razonable y de un proceso de ingeniería de variables bien estructurado, es posible construir modelos de riesgo de crédito que sean simultáneamente **útiles para el negocio** y **rigorosos desde el punto de vista estadístico**.

9 Referencias

1. Home Credit Group. (2018). *Home Credit Default Risk*. Kaggle Competition. https://www.kaggle.com/competitions/ho credit-default-risk
2. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.
3. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD*.
4. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.
5. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2nd ed.). Springer.
6. Battagliola, M. L. (2025). *Notas Estadística Aplícada III*. ITAM - Estadística Aplicada III.