

LendSmart Credit Risk Analysis

Adrián Tavera Aquino, A01659113 Daniela Robles Estrada, A01659074
Elian Alejandro López de Alba, A01659582

11 noviembre 2025

1. Problema de negocio

LendSmart es una fintech dedicada a otorgar créditos personales y a pequeños negocios. Al analizar la información histórica de su cartera, confirmamos que aproximadamente un 28% de los préstamos termina en incumplimiento. Este nivel de riesgo es demasiado alto pues presiona la rentabilidad, incrementa la necesidad de provisiones y limita la capacidad de la empresa para crecer de manera adecuada.

Frente a este contexto, se construyó un modelo estadístico de clasificación que ayude a detectar, desde la etapa de solicitud, qué clientes tienen mayor probabilidad de caer en default. Primeramente, se realizó una inspección inicial de la base de datos para asegurar que estuviera completa y bien estructurada (tipos de datos, valores faltantes, estadísticas descriptivas y proporción de préstamos en incumplimiento). Una vez validada la calidad de la información, se desarrollaron y compararon dos modelos: El Análisis Discriminante Lineal (LDA) y Análisis Discriminante Cuadrático (QDA), con el objetivo de recomendarle a LendSmart una herramienta que no sólo fuera precisa, sino también estable y fácil de explicar.

2. Hallazgos clave e insights

El análisis exploratorio confirmó que el incumplimiento no es un evento aleatorio, sino que distinguen a los clientes que son buenos pagadores de aquellos que caen en default. Al examinar la variable objetivo, se encontró una tasa de incumplimiento del 26.56%, cifra que se alinea con el 28% de demora planteado en el contexto inicial y que sirvió como referencia clave para evaluar la eficacia de los modelos.

Al comparar, mediante diagramas de caja, distintas variables financieras contra el estado del préstamo, se observó que el puntaje crediticio y el historial de pago son factores centrales. Las personas que no caen en default suelen tener un puntaje crediticio más alto y un comportamiento de pago previo más sólido. En cambio, los clientes que sí incumplen se concentran en valores de calificación crediticia claramente peores.

También se observó que la razón deuda-ingreso y la utilización del crédito son señales de riesgo, pues quienes ya destinan una proporción elevada de su ingreso al pago de deudas, o utilizan casi todo el crédito disponible en sus líneas, aparecen con mucha mayor frecuencia en el grupo de default, es decir, las personas sobreendeudadas tienen muchas más probabilidades de fallar en nuevos compromisos de pago.

La capacidad de ahorro y el patrimonio también marcan diferencia. Los clientes que ahorran más y cuentan con mayor valor de bienes, por lo que tienen un “colchón” financiero para enfrentar imprevistos; estos perfiles aparecen con menor probabilidad de incumplimiento. Por el contrario, quienes casi no ahorran y tienen pocos bienes son más vulnerables a no pagar.

El ingreso anual, los años de empleo y la estabilidad laboral son variables que completan el análisis. A mayor ingreso, más años trabajando y mayor estabilidad en el empleo, menor es la probabilidad de caer en default, por lo que un flujo de ingresos estable facilita cumplir con los pagos.

En el análisis de variables categóricas se detectaron otros adicionales, aunque secundarios frente a las variables financieras. Los clientes con niveles educativos más altos presentan tasas de incumplimiento más bajas. Algo similar ocurre con el estado civil las personas casadas tienden a incumplir menos que las solteras, divorciadas o viudas.

Asimismo con ayuda de una matriz de correlación, se puede describir, que el perfil de un solicitante de alto riesgo para LendSmart radica en personas con puntaje crediticio y comportamiento de pago previos débiles, alta razón deuda-ingreso, alta utilización de crédito, poca capacidad de ahorro, bajos niveles de activos, ingreso limitado y menor estabilidad laboral.

3. Desempeño del modelo y selección

Los modelos LDA y QDA asumen que, dentro de cada clase (clientes en default y clientes sin default), las variables predictoras siguen aproximadamente una distribución normal multivariada. Aunque en la práctica, los datos financieros rara vez cumplen esta condición de forma perfecta, en la exploración previa se observaron distribuciones regulares y sin valores extremos, por lo que se consideró que la aproximación de normalidad es

aceptable. Además, estos métodos suelen ser robustos siempre que las desviaciones no sean severas.

La diferencia clave entre LDA y QDA está en cómo tratan la matriz de covarianza de cada clase. LDA asume que las dos clases comparten la misma matriz de covarianza haciendo que el modelo sea más sencillo, más estable e interpretable. Por el contrario, QDA no impone esta restricción y permite que cada clase tenga su propia matriz de covarianza dándole al modelo mayor flexibilidad para adaptarse a situaciones en las que la variabilidad financiera de los clientes que incumplen y los que no incumplen es realmente distinta, y le permite generar fronteras de decisión curvas o cuadráticas.

Con base en esta teoría, la hipótesis inicial era que, en un contexto de riesgo de crédito, los clientes en default y los que pagan bien podrían presentar patrones de variabilidad diferentes en variables como ingreso, calificación crediticia o ahorro. Si esto fuera así, QDA podría captar mejor la estructura real de los datos y ofrecer un desempeño ligeramente superior al de LDA. Sin embargo, esta ventaja potencial debía contrastarse con evidencia empírica.

Después de preparar los datos, se entrenaron ambos modelos y se evaluaron con un conjunto de prueba independiente. Con esto, tanto LDA como QDA clasificaron correctamente todos los casos del conjunto de prueba. Los 378 clientes sin default fueron predichos correctamente como no default, y los 122 clientes restantes en default también fueron identificados correctamente como tal. Las matrices de confusión de ambos modelos son perfectas, sin falsos positivos ni falsos negativos. Técnicamente, esto se traduce en medidas de precisión y recall iguales a 1.0 para ambas clases y en curvas ROC con un área bajo la curva de 1.0.

El hecho de que los dos modelos logren un desempeño idéntico y perfecto en la muestra de prueba indica que, con la información disponible, las clases están muy bien separadas y que la flexibilidad adicional de QDA no aporta una mejora visible. En cambio, sí introduce una estructura más compleja que, en un entorno real y cambiante, podría ser más sensible al sobreajuste. Por esta razón, dado que LDA ofrece el mismo nivel de certeza y con una formulación más simple, se decidió seleccionar el modelo LDA como la opción más fácil de interpretar para los datos de LendSmart. Cada variable tiene un coeficiente lineal cuyo signo y magnitud, como: un mejor historial de pago y mayor estabilidad laboral reducen fuertemente la probabilidad de default, mientras que una alta utilización del crédito y una razón deuda-ingreso elevada la incrementan.

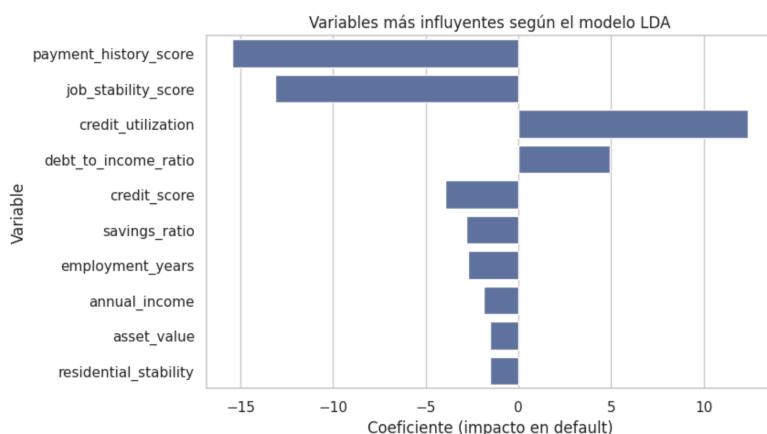


Figure 1: Variables más influyentes según el modelo LDA. Coeficientes negativos reducen la probabilidad de default, mientras que los positivos la aumentan.

4. Recomendación final

A partir de los resultados de nuestra recomendación es implementar el modelo LDA como apoyo en el proceso de originación de créditos, iniciando con una fase de prueba.

Desde la perspectiva de negocio, el beneficio principal es la posibilidad de reducir de forma importante las pérdidas por incumplimiento al identificar de forma temprana a los solicitantes con mayor riesgo. Al mismo tiempo, el modelo permite que no se terminen castigando a clientes que, por sus características financieras, sí son buenos candidatos para recibir crédito.

En la práctica, es normal que al implementar el modelo surjan errores, ya sea marcando a buenos clientes como riesgosos o no detectando a quienes no pagarán; por ello, LendSmart deberá definir un umbral que reduzca los impagos sin rechazar a demasiados solicitantes válidos. Para mejorar esto, se propone realizar una prueba piloto de varios meses donde el modelo funcione en paralelo a las políticas actuales. Esto permitirá monitorear las tasas reales de aprobación e incumplimiento, ajustar el nivel de riesgo y corregir posibles sesgos. Si el desempeño en esta fase confirma el análisis, el modelo se convertirá en una herramienta clave para hacer crecer la cartera manteniendo el riesgo bajo control.