Modelos Predictivos para Detectar el Comportamiento Bancario de un Cliente

Angel Llanos Herrera y Felipe Neira Rojas

23 de Octubre del 2025

1 Contexto

El scoring, en general, en econometría, es un conjunto de métodos estadísticos y/o modelos que asignan a cada solicitante o cliente sujeto a crédito un puntaje de riesgo. Ese puntaje representa la probabilidad de observar un comportamiento bueno o malo en una ventana futura (por ejemplo, morosidad, abandono o deudas al día). Su objetivo práctico es permitir decisiones consistentes y escalables: aprobar o rechazar un crédito, fijar límites o tasas de interés, priorizar campañas o gestionar cobranzas.

En la práctica, el proceso parte desde datos históricos (sociodemográficos, bancarios, renta, deudas, entre otros) y una etiqueta de desempeño (por ejemplo, bueno o malo). Con estas bases se entrena un modelo de clasificación que estima la probabilidad de comportamiento a partir de un conjunto de variables predictoras.

En este trabajo se dispone de una base de datos de tipo *scoring*, con 21 520 observaciones de clientes y 29 variables. La descripción de cada una de ellas se presenta en la Tabla 1.

Variable	**Descripción**
ID_CLIENTE	Número identificador del cliente.
RENTA	Renta mensual promedio del cliente.
Comportamiento	Evaluación del comportamiento del cliente. *Malo*: si presenta deuda morosa, vencida o castigada durante la ventana de desempeño; *Bueno*: en caso contrario.
SEXO	Sexo del cliente. Codificación: 1 = Femenino, 0 = Masculino.
EDAD	Edad del cliente (años).
TIPO_NAC	Tipo de nacionalidad del cliente.
EST_CIVIL	Estado civil del cliente.
GSE	Nivel socioeconómico del cliente.
NCC	Número de relaciones de cuentas corrientes.
VER_TRA	Verificación de trabajo en los últimos 2 años.
AVALUO	Avalúo de bienes raíces.
NDIR	Número de direcciones del cliente.
VER_DOM	Verificación de domicilio en los últimos 2 años.
NACR	Número de acreedores (último mes).
LINEACRNU	Línea de crédito no utilizada (último mes).
DDA_HIPOT	Deuda hipotecaria (último mes).
NUM_VIG_12	Número de meses (últimos 12) con deuda vigente.
VC_CAS	Indicador de 1+ meses (últimos 12) con deuda vencida o castigada.
IND_BC	Indicador del boletín comercial en los últimos 6 meses.
NUM_MOR_12	Número de meses (últimos 12) con deuda morosa.
MAXUTIL	Indicador de máxima utilización de línea de crédito.
RPXLCNU	Relación de línea de crédito no utilizada (promedio a máximo).
RPXDCOM	Relación de deuda comercial (promedio a máximo).
RPXDCON	Relación de deuda de consumo (promedio a máximo).

Variable	Descripción
IND_IF	Indicador del informe financiero en los últimos 6 meses.
TMOT_06	Monto total en instituciones financieras (U.F.) en los últimos 6 meses.
TDOC	Total de documentos en el boletín comercial.
TMOT_12	Monto total en instituciones financieras (U.F.) en los últimos 12 meses.
NTCRED	Número de tarjetas de crédito en los últimos 6 meses.

Diccionario de variables de la base de datos bancaria

A modo general, la forma en que se trabajará será: analizar la calidad de los datos, tratar valores faltantes, análisis exploratorio de los datos, para luego llegar al ajuste del modelo de clasificación mediante regresión logística para el comportamiento del cliente. Finalmente, se evaluará un modelo de clasificación para el comportamiento crediticio de 5 diferentes perfiles de clientes para identificar la probabilidad de mal comportamiento crediticio.

Es importante mencionar que de forma aleatoria, se eliminan 100 observaciones.

2 Calidad de los datos

La calidad de los datos es el primer control necesario para los modelos de scoring: determinando la confiabilidad de las conclusiones y ajustes. Es decir, afectan directamente a las decisiones de aprobación, tasas de interes y limites. Una base de datos con errores, valores faltantes e inconsistencias puede disminuir la discriminación y sesgar el balance. Por eso, antes del modelado, revisaremos y corregiremos problemas con respecto a la calidad de los datos.

2.1 Valores faltantes

En un principio, se detectaron dentro de todas las variables todos los registros no nulos. Sin embargo, es necesario identificar valores faltantes dados por errores de tabulación (valores tabulados como NA o variantes). Por lo que se revisarán las categorías de las variables categóricas. Donde, se encontraron en la variable grupo socioeconómico (GSE) las siguientes categorías.

GSE	Conteo	Porcentaje
NA	13,550	63.26%
C3	2,680	12.51%
D	2,135	9.97%
C2	1,780	8.31%
AB	981	4.58%
E	293	1.37%
nA	1	0.00%

Table 2: Distribución de *GSE*: conteo y porcentaje

En cuanto a GSE, podemos notar la existencia de dos categorías; NA con 13550 observaciones representando el 62.26% de las observaciones. También, 1 registro nA, el cual también representa un valor faltante. Estas categorías se redefinen como valores faltantes (NA). Entonces, dentro de todas las variables, es en GSE donde se encuentran los valores faltantes, representando más de la mitad de las observaciones totales.

Ahora, será necesario revisar el tratamiento de estos valores faltantes. Sin antes mencionar que plantear la "imputación" y estimación de estas categorías de GSE dentro del contexto no es del todo valido, porque GSE corresponde a una clasificación estandarizada. Donde, por ejemplo, el estándar vigente de AIM Chile establece que el GSE se calcula a partir de un índice socioeconómico construido con tres insumos específicos del hogar: (i) ingreso per cápita equivalente (que ajusta el ingreso por tamaño del hogar con una escala $n^{0.7}$), (ii) nivel educativo del principal sostenedor y (iii) ocupación del principal sostenedor. Además, el modelo se calibra con datos de CASEN y de la Encuesta de Presupuestos Familiares (EPF). En el contexto de la base de datos scoring estas variables no se encuentran presentes (tamaño del hogar, educación u ocupación del sostenedor), siendo complicado asignar GSE de forma generalizable. AIM. (AIM Chile, 2024).

En 2018 se realizó una actualización de GSE, donde, antes de esta actualización la educación y la ocupación eran los principales ejes, sin embargo, luego de 2018 dió paso a un modelo explicito basado en ingreso per capita, educación y ocupación, con siete categorías y parametros obtenidos de estadisticas publicas (CASEN y EPF). Este cambio se hizó para asegurar comparabilidad, transparencia y actualización. Intentar estimar la clasificación de GSE desde variables no contempladas en el modelo deja de ser GSE. Es por esta razon que lideres de la industria enfatizan que con la actualización del GSE se puede notar que no es linealmente comparable con la versión previa.(GfK Chile, 2019).

De todas formas, incluyendo variable/s que tengan sentido con las incluídas dentro del modelo original (a pesar de no ser la/s mismas), se compararán modelos para revisar su rendimiento.

2.2 Valores atípicos

A modo de comprensión e identificación de las variables, es importante cuantificar los valores atípicos de manera univariada bajo criterio de Tukey porque estos puntos extremos pueden distorsionar media, varianza, sesgar relaciones en el ajuste de modelo y afectar la detección de patrones.

Revisando un ranking de los valores atípicos dentro de las variables.

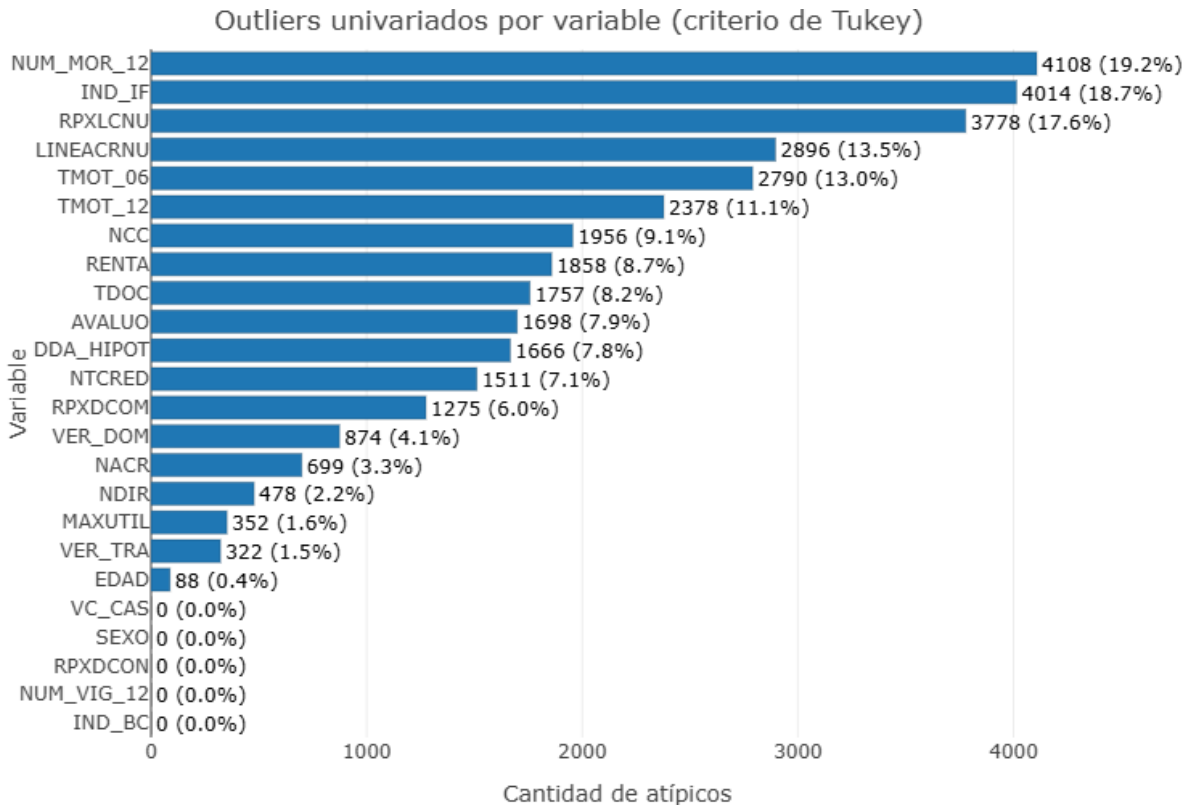


Figure 1: Ranking de variables con valores atípicos

Se detectan un gran numero de valores atípicos en la variable "NUM_MOR_12", con 4108 (19.2%) casos atípicos. En la variable "IND_IF" se detectan 4014 (18.7%), y luego en "RPXLCNU" con 3778 (17.6%). Estas son las tres principales variables con mayor cantidad de valores atípicos.

2.3 Tratamiento de valores faltantes en grupo socioeconómico (GSE)

Habiendo visto que las variables que componen la categorización de GSE no se encuentran en la base de datos, pero entendiendo que la única que más se acerca es RENTA, o bien, la renta mensual del cliente, esta sería la única variable que nos aportaría información valiosa a la clasificación. Sin embargo, se probará si existen diferencias y asociaciones importantes con otras variables.

Luego de esto, se aplicará un modelo de regresión multilogística para validar el rendimiento sobre la imputación de valores faltantes en el grupo socioeconómico (GSE), si este rendimiento es óptimo en la validación (¿ 70% de precisión en cada una de las categorías) se utilizaría para la imputación de estos valores. Además, en un estudio aplicado a credit scoring, se compararon imputaciones por media/mediana versus KNN, concluyendo que KNN preserva mejor las relaciones entre variables y mejora el rendimiento del modelo (Parveen & Thangaraju, 2024).

2.3.1 Selección de variables

Para la selección de variables con enfoque en GSE (AB, C3, C2, D, E) partimos con que esta etiqueta refleja rasgos socioeconómicos que suelen vincularse con renta, avaluó y otras características. Por lo que se revisarán estas asociaciones y patrones.

Primero analizamos la asociación entre variables categóricas y GSE mediante la prueba de chi-cuadrado de independencia; cuando se observaron asociaciones significativas, estimamos la V de Cramer para valorar la magnitud del efecto. Con muestras grandes, la significancia estadística puede aparecer aun con efectos pequeños; por ello, utilizamos la V de Cramer como criterio práctico y consideramos relevantes solo las asociaciones con tamaño de efecto al menos modesto.

Variable	**p-value**	**V de Cramer**
NUM_VIG_12	9.005420e−07	0.059
SEXO	3.904453e−04	0.051
NACR	8.857879e−03	0.032
NUM_MOR_12	1.107323e−02	0.048
VC_CAS	2.755653e−02	0.037
Comportamiento	3.113453e−02	0.037

Table 3: Asociación con GSE: prueba Chi-cuadrado (p-value) y tamaño de efecto (V de Cramer).

Los resultados muestran que varias categóricas presentan diferencias significativas entre niveles de GSE, pero con tamaños de efecto débiles (V de Cramer ¡ 0.10), por lo que su aporte práctico a la discriminación es limitado.

En paralelo, para las variables numéricas contrastamos diferencias entre niveles de GSE con Kruskal–Wallis, complementando la significancia con la inspección de la magnitud de las diferencias.

Variable	**p-value**	ε^2	**Efecto**
RENTA	3.94e−317	0.186	grande ($\geq$ 0.14)
AVALUO	1.26e−265	0.156	grande ($\geq$ 0.14)
NCC	4.11e−47	0.028	pequeño (aprox 0.01)
RPXLCNU	4.26e−13	0.008	$<$ 0.01 (muy pequeño)

Table 4: Top 4 por ε^2 (Kruskal–Wallis): significancia y tamaño de efecto.

Entre las numéricas, RENTA y AVALUO destacan por su asociación fuerte y consistente con GSE.Ahora, dado que ambas están altamente correlacionadas (correlación de 0.86) entre sí, mantenerlas simultáneamente introduciría redundancia y potencial multicolinealidad, dificultando la asignación estable de pesos y pudiendo degradar el rendimiento. Por ese motivo priorizamos RENTA como predictor principal y excluimos AVALUO, privilegiando un modelo más parsimonioso, interpretable y robusto. Finalmente, las categóricas con efecto débil no se incorporaron, aun cuando mostraran significancia, porque su impacto práctico sobre la predicción de GSE es marginal.

2.3.2 Ajuste de modelo de regresión multilogística

Ajustando el modelo de regresión multilogística para la variable dependiente "GSE" (sin considerar NA), con la variable independiente de "RENTA", dividiendo en 80% en entrenamiento y el 20% para validación, nos entrega la siguiente matriz de confusión.

	Real				
Predicción	**AB**	**C2**	**C3**	**D**	**E**
AB	62.82%	17.31%	19.87%	0.00%	0.00%
C2	0.00%	0.00%	0.00%	0.00%	0.00%
C3	11.93%	25.51%	36.42%	24.24%	1.90%
D	0.64%	20.35%	34.66%	37.52%	6.84%
E	0.00%	0.00%	0.00%	0.00%	0.00%

Table 5: Matriz de precisión por clase predicha regresión multilogística.

Viendo esto, podemos notar que la precisión del modelo de regresión multilogística en cada una de las categorías del grupo socioeconómico es de 62.82% en AB, 0% para C2, 36.42% para C3, 37.52% para D y 0% en la categoría E. Por lo tanto, este rendimiento se considera pobre e inutilizable de manera válida.

2.3.3 Ajuste de modelo K-Nearest Neighbors (KNN)

Ahora utilizando KNN para la clasificación de GSE y utilizando renta como predictora. Además, se seleccionará los k-vecinos el cual optimice la métrica accuracy.

El numero de k-vecinos óptimos para el accuracy es 31, con un accuracy de 0.477, el cual igualmente se considera con un rendimiento pobre.

Ajustando el modelo KNN con k = 31 para la variable dependiente "GSE" (sin considerar NA), con la variable independiente de "RENTA", dividiendo en 80% en entrenamiento y el 20% para validación, nos entrega la siguiente matriz de confusión.

	Real				
Predicción	**AB**	**C2**	**C3**	**D**	**E**
AB	62.80%	15.46%	14.01%	6.76%	0.97%
C2	8.43%	38.20%	30.34%	19.66%	3.37%
C3	2.41%	25.88%	53.38%	16.56%	1.77%
D	6.36%	16.78%	21.38%	48.59%	6.89%
E	0.00%	0.00%	0.00%	0.00%	0.00%

Table 6: Matriz de precisión por clase predicha (porcentaje por fila).

Podemos notar que en varios casos la probabilidad de acierto (precisión) de la clasificación sube al menos un 10% en comparación al modelo de clasificación multilogístico. Para la categoría AB se presenta una diferencia insignificante con el modelo de multilogístico de 62.8%, en la categoría C2 aumenta a 38.2%, luego de 53.38% para C3, para D es de 48.59% y nuevamente de 0% para la categoría de E, afectado por el desbalance en esta categoría.

2.3.4 Decisión sobre variable de grupo socioeconómico (GSE)

Cuando la fracción de datos faltantes en una variable es pequeña, imputar (rellenar) esos vacíos suele no distorsionar las conclusiones: las distribuciones, las medias y las relaciones con otras variables se conservan razonablemente bien. En cambio, a medida que el porcentaje de ausencia crece, y en especial cuando los datos son "completamente al azar" (no-MCAR), por ejemplo, cuando faltan más justo en los casos con ingresos muy altos o muy bajos, la imputación puede introducir sesgos. En esos escenarios, el método "adivina" sistemáticamente valores con menor información y termina empujando el análisis hacia estimaciones demasiado optimistas o conservadoras.

Un artículo de Romero-Duque et al, sugiere una regla práctica: antes de imputar, reducir el faltante total a alrededor de 10% o menos. Eliminando variables que concentran la mayor parte de los vacíos, siempre que su eliminación no comprometa la validez del estudio. Con ese recorte previo, la imputación trabaja sobre "huecos" residuales más pequeños y menos estructurados, lo que disminuye el riesgo de sesgo y ayuda a preservar las relaciones reales entre variables. (Romero-Duque et al., 2023).

Como ninguno de estos modelos logra clasificar GSE con un rendimiento y robustez optimo, utilizar estos modelos y variables no resulta valido al presentar más de un 50% de valores faltantes dentro de su variable, por

lo tanto, se estaría derivando en errores y sesgos altos. Esta variable al no ser útil tal como está, y tampoco valida su estimación, se eliminará del análisis.

3 Análisis exploratorio de los datos

3.1 Univariado

Dentro del análisis univariado se consideraron el análisis de las variables numéricas y categóricas más importantes, siendo las siguientes:

3.1.1 Comportamiento

Comportamiento	Porcentaje
Malo	52.33%
Bueno	47.67%

Table 7: Frecuencia Relativa Porcentual del Comportamiento del Cliente

Primeramente consideramos analizar la variable objetivo para nuestro trabajo predictivo, la cual refiere al comportamiento bancario del cliente. Podemos notar que predomina levemente la categoría malo por sobre el bueno, con un 52,33% de clientes que pertenecen a esta categoría, mostrando una mayor tendencia a comportamientos malos.

3.1.2 Renta

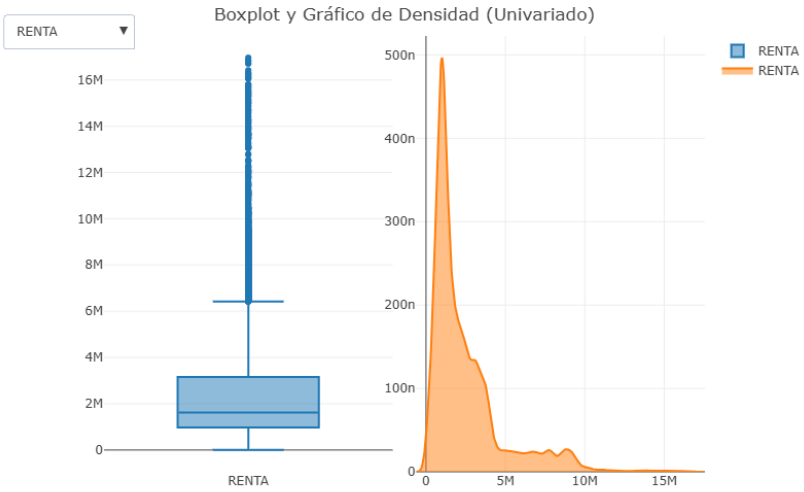


Figure 2: Gráfico Distribucional de la Renta de los Clientes

Estadístico	RENTAS
Mínimo	0
1er Cuartil	973 862
Mediana	1 615 486
Media	2 484 871
3er Cuartil	3 154 774
Máximo	16 960 791

Table 8: Resumen Descriptivo de la Variable

Es posible ver que al menos el 50% de los clientes cuentan con una renta menor o igual a 1.62 millones de pesos (M) y al menos un 25% tienen una renta mayor o igual a 3.16 M, con una mediana de 1.62 M y un rango entre 0 y 16.96 M. Además, la media (2.49 M) es mayor que la mediana (1.62 M), lo que sugiere valores atípicos altos, notorio en la cola derecha del gráfico de densidad.

3.1.3 Sexo

Comportamiento	Porcentaje
Masculino	51.53%
Femenino	48.47%

Table 9: Frecuencia Relativa Porcentual del Sexo del Cliente

Analizando el sexo de los clientes, notamis nuevamente categorias balanceadas, en dónde el sexo masculino predomina levemente con un 51,53% del total de clientes.

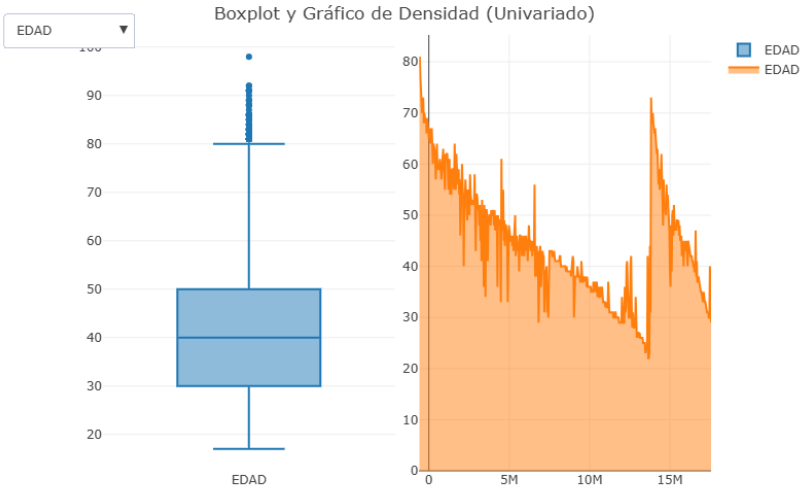
3.1.4 Estado Civil

Categoría	Porcentaje
Casado (C)	53.08%
Soltero (S)	41.25%
Divorciado (D)	2.97%
Viudo (V)	1.56%
No informado (N)	1.13%

Table 10: Distribución Porcentual de la Variable Estado Civil

La variable Estado Civil muestra un predominio de la categoría Casado (53.08%), seguida por Soltero (41.25%), mientras que las demás categorías —Divorciado (2.97%), Viudo (1.56%) y No informado (1.13%)— representan proporciones mucho menores.
Esto indica una población principalmente casada o soltera, con una distribución relativamente equilibrada entre ambos grupos principales, y una baja representación de las demás condiciones civiles.

3.1.5 Edad



Estadístico	EDAD
Mínimo	17.00
1er Cuartil	30.00
Mediana	40.00
Media	41.07
3er Cuartil	50.00
Máximo	98.00

Table 11: Resumen Descriptivo de la Variable

Figure 3: Gráfico Distribucional de la Edad de los Clientes

Es posible ver que al menos el 50% de los clientes tienen una edad menor o igual a 40 años y al menos un 25% tienen una edad mayor o igual a 50 años, con una mediana de 40 años y un rango entre 17 y 98 años. Además, la media (41.08) es mayor que la mediana (40), lo que sugiere ligera cola derecha (edades altas). Podemos notar también una gran cantidad de clientes en un rango pequeño de edad (el 25% de los clientes tiene entre 30 y 40 años).

3.1.6 Avalúo

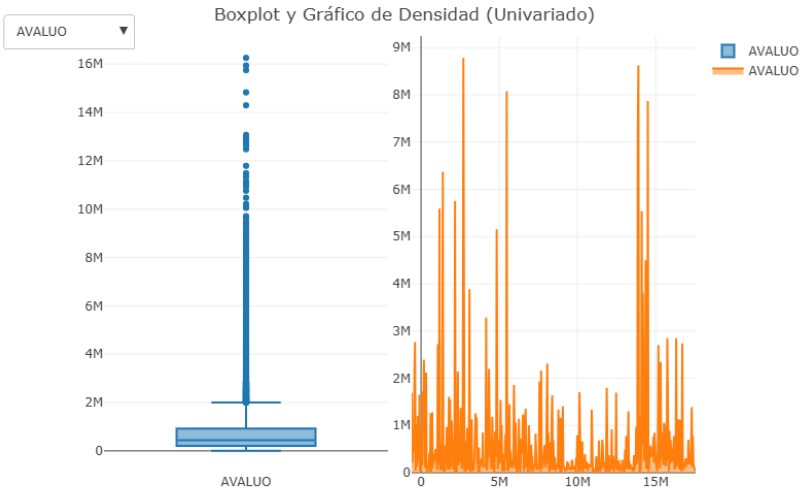


Figure 4: Gráfico Distribucional del Avalúo

Estadístico	AVALÚO
Mínimo	0
1er Cuartil	203 041
Mediana	443 408
Media	763 838
3er Cuartil	922 936
Máximo	16 260 850

Table 12: Resumen Descriptivo de la Variable

Es posible ver que al menos el 50% de los clientes cuentan con un avalúo menor o igual a 0.44 M y al menos un 25% tienen un avalúo mayor o igual a 0.92 M, con una mediana de 0.44 M y un rango entre 0 y 16.26 M. Además, la media (0.76 M) es mayor que la mediana (0.44 M), lo que sugiere atípicos altos/cola derecha. En este caso se nota una distribución similar a la variable Renta, con un distinto rango de valores

3.1.7 Cantidad de Cuentas Corrientes

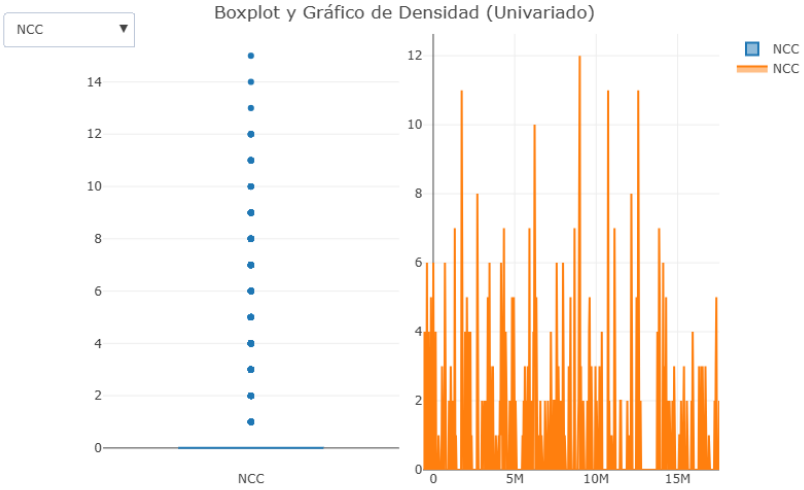


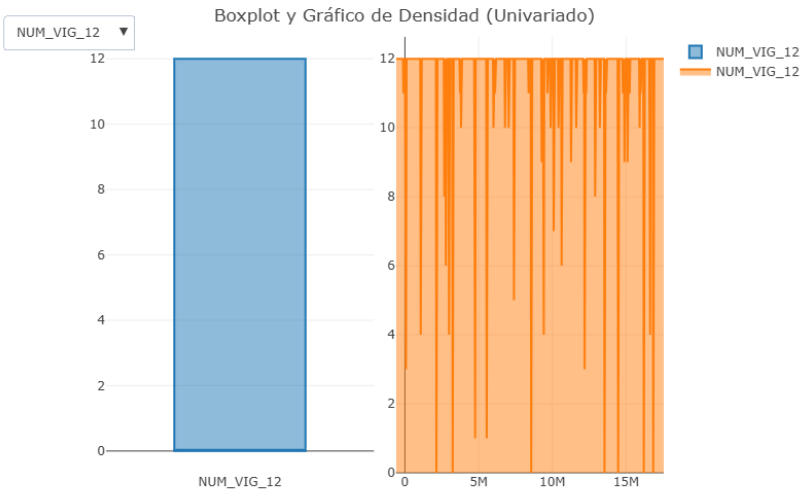
Figure 5: Gráfico Distribucional del Número de Cuentas Corrientes

Estadístico	NCC
Mínimo	0.0000
1er Cuartil	0.0000
Mediana	0.0000
Media	0.2496
3er Cuartil	0.0000
Máximo	15.0000

Table 13: Resumen Descriptivo de la Variable

Es posible ver que al menos el 75% de los clientes cuentan con 0 cuentas corrientes, mismo valor de su mediana y un rango entre 0 y 15. Esta variable muestra como valores atípicos a personas que cuentan con 1 o más cuenta corrientes, mostrando técnicamente sesgos en valores altos.

3.1.8 Deuda vigente los últimos 12 meses



Estadístico	NUM_VIG_12
Mínimo	0.000
1er Cuartil	0.000
Mediana	0.000
Media	3.678
3er Cuartil	12.000
Máximo	12.000

Table 14: Resumen Descriptivo de la Variable

Figure 6: Gráfico Distribucional del Número de Meses con Deuda Vigente

Es posible ver que al menos el 50% de los clientes cuentan con 0 meses con deuda vigente y al menos un 25% tienen exactamente 12 meses, con una mediana de 0 y un rango entre 0 y 12. Además, la media (3.68) es mayor que la mediana (0), lo que sugiere cola derecha (subgrupo con muchos meses vigentes). Este comportamiento puede mostrar principalmente casos demasiado extremos, en donde los clientes no tienen deuda vigente, o suelen pagar su deuda el último mes.

3.1.9 Variables Categóricas Desbalanceadas

Se observa un marcado predominio de clientes chilenos dentro del tipo de nacionalidad (alrededor de 98%), mientras que las categorías minoritarias, correspondientes a personas naturalizadas y extranjeras, representan proporciones cercanas al 1% cada una. También se aprecian tasas muy bajas de verificación reciente, tanto en trabajo como en domicilio: la verificación de trabajo en los últimos dos años alcanza cerca de 98.5% y la verificación de domicilio alrededor de 96%, quedando la gran mayoría con verificación registrada. En la carga crediticia del último mes, la mayor parte de los clientes no presenta acreedores, y los niveles superiores son poco frecuentes. Asimismo, el indicador de máxima utilización de la línea de crédito aparece prácticamente en toda la muestra, con escasos casos sin esa condición.

Con un desbalance menos extremo, el antecedente de haber tenido meses con deuda vencida o castigada en el último año se distribuye aproximadamente en 70% sin eventos y 30% con al menos un mes afectado. Los registros de comportamiento en instituciones financieras y en boletín comercial durante los últimos seis meses muestran mayor proporción de clientes sin eventos, aunque con grupos no menores que sí han tenido registros. Finalmente, el número de tarjetas de crédito emitidas en los últimos seis meses se concentra fuertemente en cero, con una fracción pequeña con una tarjeta y porcentajes marginales en valores superiores.

3.2 Bivariado

Para el análisis bivariado, se consideraron estudiar relaciones de diversas variables, tanto numéricas como categóricas, con la variables de interés (Comportamiento). Todo esto con el fin de identificar si existe alguna tendencia en alguna variable con respecto a los diversos comportamientos.

3.2.1 Renta - Comportamiento

Los clientes con comportamiento bueno muestran una distribución más concentrada en rentas bajas-medias, mientras que los malo presentan una cola más extensa hacia rentas altas, Esto no indica una relación directa entre el comportamiento del cliente y su renta, pero si se reporta una mayor cantidad de casos de comportamiento malo que de bueno cuando la renta es mayor. Por otra parte, se nota una mayor cantidad de comportamiento Bueno cuando la renta es menor al millón de pesos

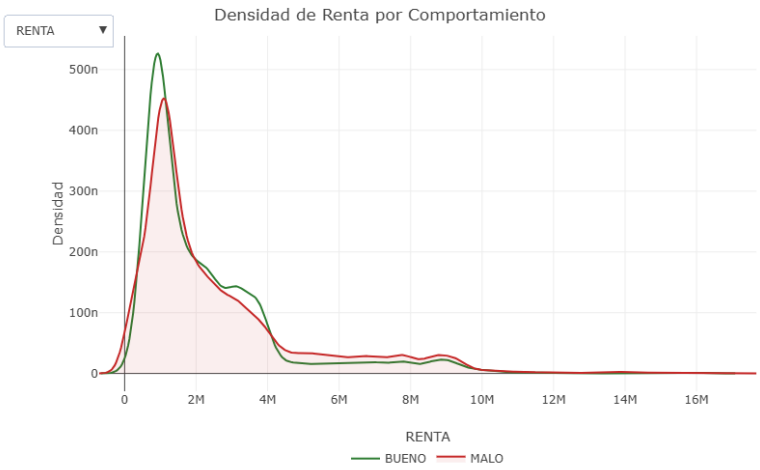


Figure 7: Gráfico Distribucional de la Renta según Comportamiento

3.2.2 Edad - Comportamiento

Los clientes con comportamiento malo se concentran entre los 30 y 50 años, rango donde también hay mayor densidad de comportamiento bueno, aunque la categoría Malo con mayor predominio, lo que sugiere que la edad media adulta presenta mayor riesgo crediticio, pero no con una tendencia muy notoria.

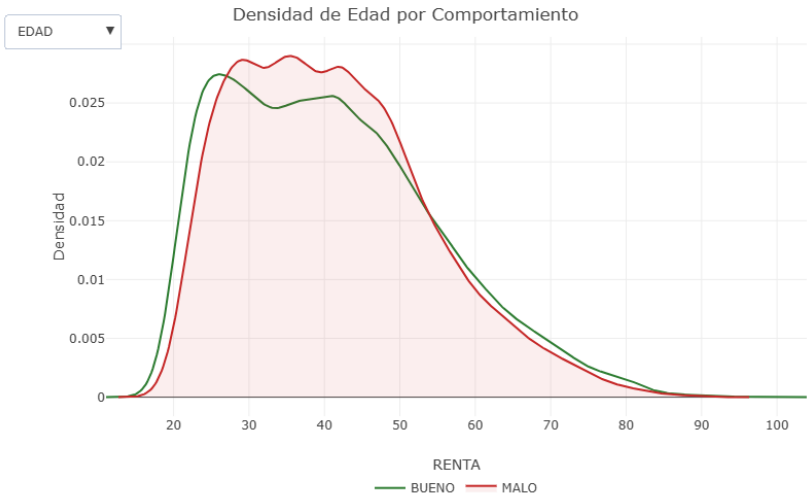


Figure 8: Gráfico Distribucional de la Edad según Comportamiento

3.2.3 Avalúo - Comportamiento

Ambos grupos se concentran en avalúos bajos, pero los clientes de comportamiento malo presentan densidad ligeramente mayor en valores altos, lo que sugiere que tener bienes de alto avalúo no necesariamente se asocia a mejor comportamiento crediticio. Esta comparación tiene una interpretación similar a la anterior. Nuevamente avalúos bajos pueden mostrar mayor cantidad de comportamiento bueno.

Figure 9: Gráfico Distribucional deL Avalúo según Comportamiento

3.2.4 Deuda Hipotecaria - Comportamiento

Los clientes de comportamiento malo exhiben una distribución más amplia hacia valores positivos, con mayor dispersión de deuda hipotecaria, mientras que los clientes de buen comportamiento se agrupan cerca de cero, indicando niveles más controlados de endeudamiento.

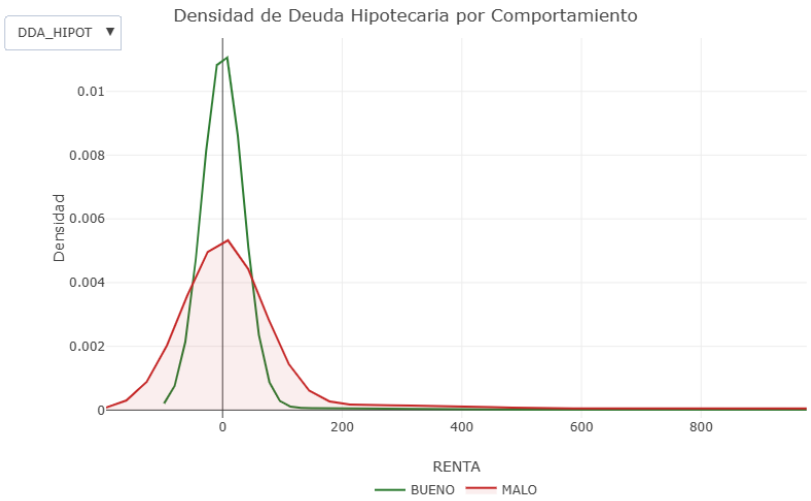


Figure 10: Gráfico Distribucional de la Deuda Hipotecaria según Comportamiento

3.2.5 Sexo - Comportamiento

La proporción de clientes con comportamiento bueno es ligeramente mayor en mujeres que en hombres, aunque en ambos casos los perfiles malo también presentan una fracción relevante. En cuanto al comportamiento de los hombres, es predominante la categoría malo con un 55% de observaciones.

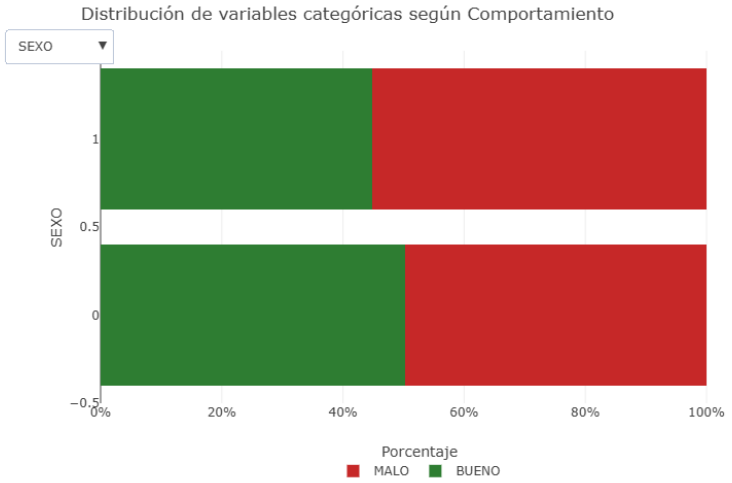


Figure 11: Gráfico Distribucional del Sexo del cliente según Comportamiento

3.2.6 Tipo de Nacionalidad - Comportamiento

Los clientes con nacionalidad extranjera muestran un leve predominio de comportamiento bueno respecto a los nacionales y chilenos, lo que podría reflejar un perfil financiero más estable o menor exposición crediticia. Es importante recordar que hay una baja cantidad de personas fuera de la categoría chilena, por lo tanto es importante tener en cuenta principalmente la categoría Chilenos (C).

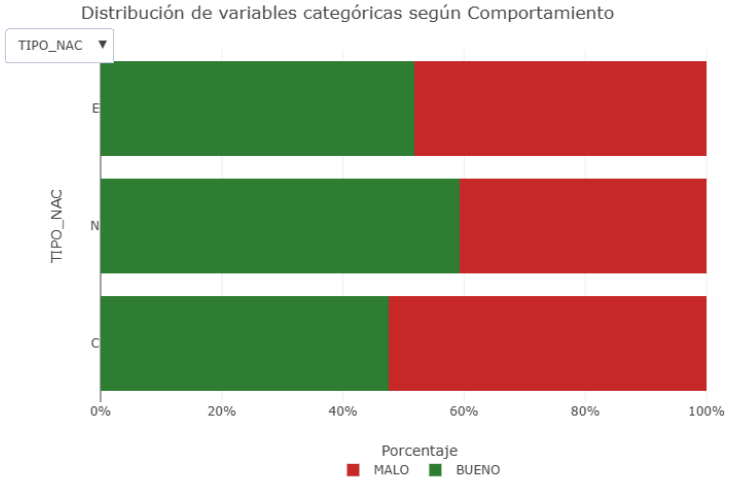


Figure 12: Gráfico Distribucional del Tipo de Nacionalidad según Comportamiento

3.2.7 Estado Civil - Comportamiento

Las categorías “Casado” ”Divorciado” y “Viudo” concentran mayor proporción de clientes con comportamiento malo que de comportamiento Bueno, mientras que los “Solteros” y “No informados” presentan proporciones más equilibradas o con ligera ventaja de los comportamientos buenos.

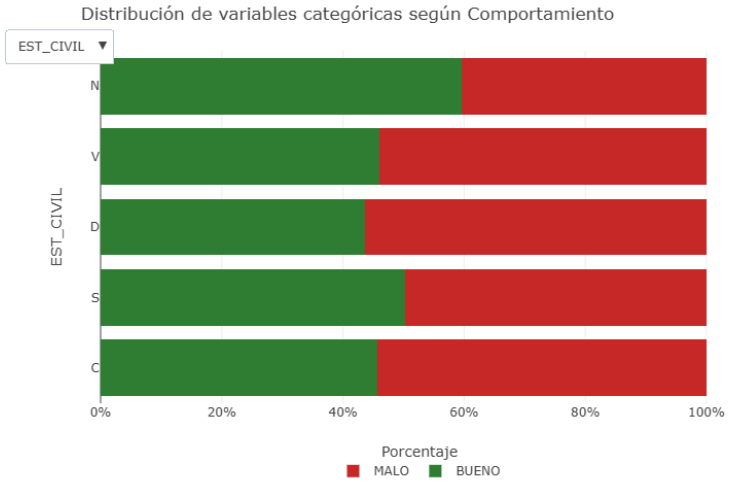


Figure 13: Gráfico Distribucional del Estado Civil según Comportamiento

3.2.8 Deuda Comercial - Comportamiento

La relación de deuda comercial promedio a máxima es más elevada y dispersa en los clientes de comportamiento malo, mientras que los de comportamiento bueno presentan valores bajos y concentrados cerca de cero. Esto refleja una utilización más constante o elevada de deuda comercial en los perfiles de riesgo.

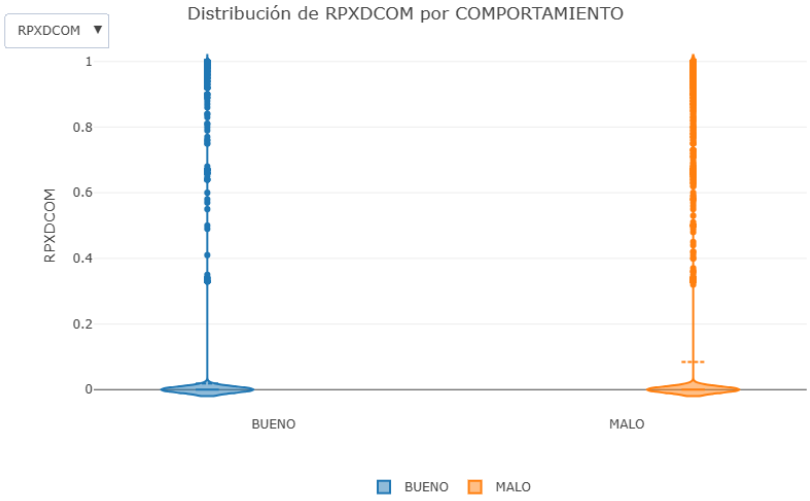


Figure 14: Gráfico Distribucional de la Deuda Comercial según Comportamiento

3.2.9 Deuda de Consumo - Comportamiento

Los clientes de comportamiento malo presentan una clara concentración en valores altos de la relación de deuda de consumo, mientras que los de comportamiento bueno mantienen proporciones cercanas a cero. Esto evidencia una mayor dependencia del crédito de consumo entre quienes muestran mal comportamiento. Esta variable tiene fuerte capacidad discriminante, muy útil para detectar el comportamiento del cliente en un posible modelo predictivo.

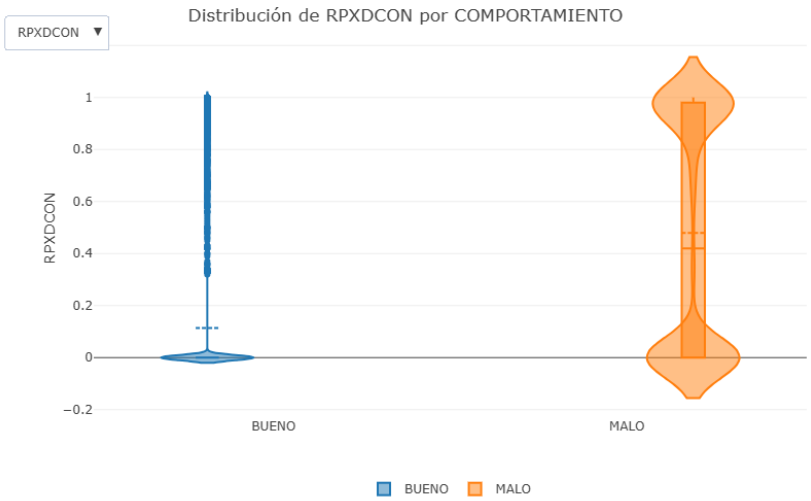


Figure 15: Gráfico Distribucional de la Deuda de Consumo según Comportamiento

3.3 Multivariado

Por último, se realiza el análisis descriptivo multivariado, que de partida nos permitirá establecer tanto relaciones entre variables, como relaciones de variables con la de interés.

3.3.1 Asociación entre Variables

Para analizar las asociaciones entre variables de distinto tipo se construyó una matriz comparativa que integra tres medidas de relación: el coeficiente de Spearman, la V de Cramér y el estadístico η^2. Cada una de ellas fue seleccionada según la naturaleza de las variables involucradas.

El coeficiente de correlación de Spearman (ρ) se utilizó para medir la fuerza y dirección de la relación entre variables numéricas u ordinales. A diferencia del coeficiente de Pearson, Spearman no asume normalidad ni linealidad estricta, siendo adecuado para distribuciones sesgadas o con valores atípicos (Conover, 1999; Field, 2013). Esta medida se basa en los rangos de los datos y captura asociaciones monótonas, tanto lineales como no lineales.

La V de Cramér se aplicó para evaluar la asociación entre variables categóricas. Derivada del estadístico chi-cuadrado, esta medida proporciona un valor normalizado entre 0 y 1 que refleja la magnitud de la relación en tablas de contingencia (Cramér, 1946; Agresti, 2019). A diferencia del chi-cuadrado tradicional, la V de Cramér corrige el efecto del tamaño de muestra y del número de categorías, permitiendo comparaciones más equilibradas entre diferentes pares de variables cualitativas.

Por último, el coeficiente η^2 (eta cuadrado) se empleó para cuantificar la proporción de varianza explicada cuando se analiza una variable numérica en función de una variable categórica. Este estadístico, común en el análisis de varianza (ANOVA), permite estimar la fuerza de asociación entre ambos tipos de variables al expresar la fracción de la variabilidad total atribuible a las diferencias entre grupos (Cohen, 1988; Olejnik & Algina, 2003).

La combinación de estas tres medidas en una matriz única facilita la visualización de asociaciones entre variables numéricas, categóricas y mixtas dentro de un mismo marco interpretativo, permitiendo identificar patrones de relación relevantes para el análisis multivariante posterior.

3.3.2 Mapa de Asociación entre Variables

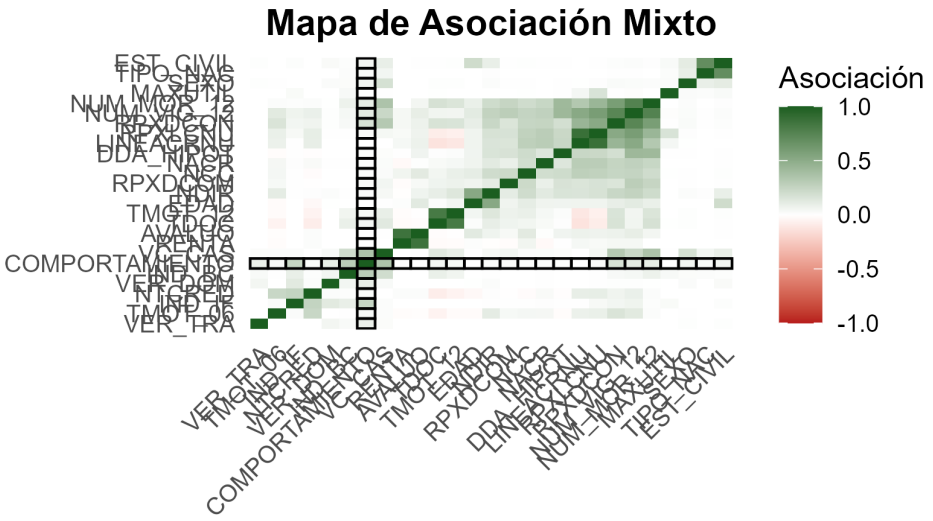


Figure 16: Mapa de Asociación entre variables

Debido a la gran cantidad de variable, visualizaremos una tabla con las asociaciones más fuertes entre variables, como también entre cada variable con la de interés (Comportamiento).

3.3.3 Asociaciones entre Variables

Las variables que tienen una relación mayor a 0.5 (ya sea positiva o negativa) entre ellas son:

Variable 1	Variable 2	Asociación
Línea de crédito no utilizada	Relación línea de crédito no utilizada prom. a máx.	0.881
Avalúo de bienes raíces	Renta mensual promedio	0.860
Total de documentos en BC	Monto total en IF (12 meses)	0.820
Meses con deuda morosa	Meses con deuda vigente	0.767
Meses con deuda vigente	Relación deuda de consumo prom. a máx.	0.741
Estado civil	Tipo de nacionalidad	0.695
Meses con deuda vigente	Relación línea de crédito no utilizada prom. a máx.	0.592
Meses con deuda morosa	Relación deuda de consumo prom. a máx.	0.592
Edad del cliente	Número de direcciones registradas	0.524
Línea de crédito no utilizada	Meses con deuda vigente	0.521
Relación deuda de consumo prom. a máx.	Relación línea de crédito no utilizada prom. a máx.	0.520

Table 15: Principales asociaciones entre variables según coeficiente de relación

Las interpretaciones son las siguientes:

- línea de crédito no utilizada en el último mes con relación de línea de crédito no utilizada promedio a máximo (Spearman = 0.881): relación muy alta; ambas capturan la misma dimensión de no utilización de la línea de crédito, por lo que pueden aportar información redundante.
- avalúo de bienes raíces con renta mensual promedio (0.860): a mayor renta se observa, en general, un mayor avalúo de propiedades.
- total de documentos en el boletín comercial con monto total en instituciones financieras en los últimos 12 meses (0.820): un mayor número de documentos se vincula a montos financieros más altos en el último año, lo que sugiere mayor actividad o intensidad crediticia.

- número de meses con deuda morosa con número de meses con deuda vigente (0.767): más meses con deuda vigente coocurren con más meses morosos, consistente con mayor exposición y mayor probabilidad de mora.
- número de meses con deuda vigente con relación de deuda de consumo promedio a máximo (0.741): mayor vigencia de deudas se asocia a una razón más elevada entre el promedio y el máximo de deuda de consumo.
- estado civil con tipo de nacionalidad (V de Cramér = 0.695): dependencia de magnitud moderada entre ambas variables categóricas.
- número de meses con deuda vigente con relación de línea de crédito no utilizada promedio a máximo (0.592): mayor vigencia de deuda se relaciona con una mayor proporción de línea de crédito no utilizada.
- número de meses con deuda morosa con relación de deuda de consumo promedio a máximo (0.592): mayor morosidad se asocia a una razón más alta de deuda de consumo.
- edad en años con número de direcciones del cliente (0.524): a mayor edad suele registrarse un mayor número de domicilios históricos, compatible con mayor trayectoria residencial.
- línea de crédito no utilizada en el último mes con número de meses con deuda vigente (0.521): más línea de crédito sin uso se observa junto con más meses con deuda vigente.

Estas relaciones orientan la selección de predictores evitando redundancias y favoreciendo un modelo predictivo parsimonioso y estable.

3.3.4 Asociaciones entre Variables con el Comportamiento del Cliente

Todas las variables tienen una asociación menor a 0.5 con la variable Comportamiento. En este caso las variables que poseen una asociación mayor a 0.1 son las siguientes:

Variable	**Asociación**	**Tipo**
Indicador de 1+ meses con deuda vencida o castigada (últimos 12 m)	0.353	cat–cat (V Cramér)
Indicador en boletín comercial (últimos 6 m)	0.298	cat–cat (V Cramér)
Indicador en instituciones financieras (últimos 6 m)	0.245	cat–cat (V Cramér)
Relación de deuda de consumo promedio a máximo	0.171	num–cat (η^2)
Meses con deuda vigente (últimos 12 m)	0.118	num–cat (η^2)
Meses con deuda morosa (últimos 12 m)	0.101	num–cat (η^2)

Table 16: Asociación de cada Variable con Comportamiento

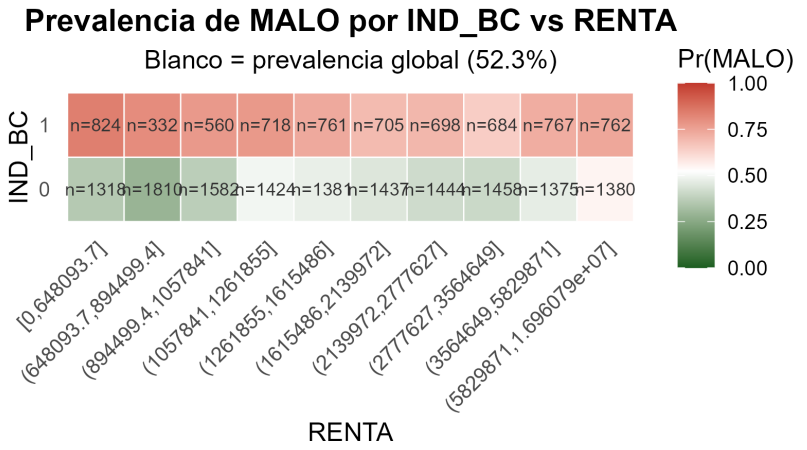
Las interpretaciones son las siguientes:

- indicador de al menos un mes con deuda vencida o castigada en los últimos 12 meses con comportamiento (V de Cramér = 0.353): tener uno o más meses con deuda vencida o castigada se asocia con mayor prevalencia de malo; dentro del grupo analizado es el indicador con mayor capacidad discriminante.
- indicador en boletín comercial en los últimos 6 meses con comportamiento (0.298): la presencia de registros en el boletín comercial se vincula a peor desempeño crediticio; el efecto es relevante y concordante con los patrones observados en los mapas de prevalencia.
- indicador en instituciones financieras en los últimos 6 meses con comportamiento (0.245): la actividad o alerta reciente en instituciones financieras se relaciona con mayor probabilidad de malo, con un tamaño de efecto moderado.
- relación de deuda de consumo promedio a máximo con comportamiento ($\eta^2 = 0.171$): a mayor razón entre el promedio y el máximo de deuda de consumo, mayor prevalencia de malo; explica una fracción apreciable de la varianza.
- número de meses con deuda vigente en los últimos 12 meses con comportamiento ($\eta^2 = 0.118$): un mayor número de meses con deuda vigente se asocia a mayor riesgo de malo, con un efecto pequeño a moderado.

- número de meses con deuda morosa en los últimos 12 meses con comportamiento ($\eta^2 = 0.101$): más meses morosos incrementan la probabilidad de malo; el efecto se ubica en el límite inferior de lo moderado.

estas relaciones ayudan a priorizar predictores en el modelado posterior, favoreciendo una selección parsimoniosa y evitando redundancias.

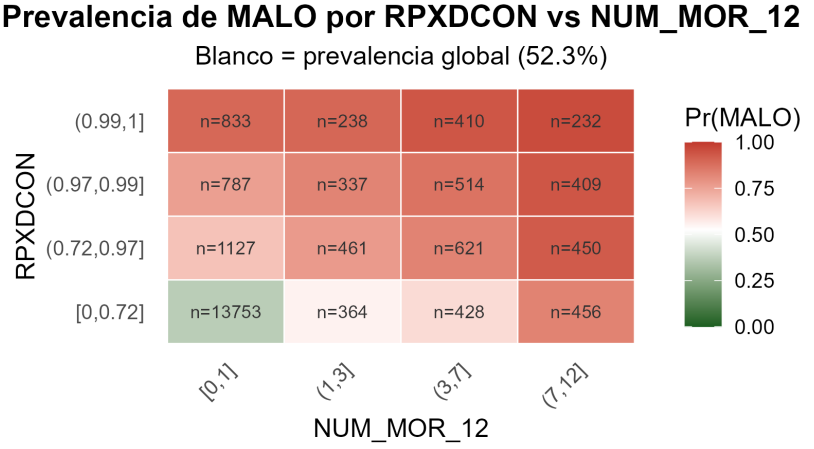
3.3.5 Prevalencia de Comportamiento Malo según Renta e Indicador BC



La prevalencia de comportamiento malo es mayor en clientes con indicador positivo en el Boletín Comercial (Indicador BC = 1) y rentas bajas a medias. En rentas altas, la prevalencia tiende a igualarse o disminuir levemente. Tener en cuenta que el Indicador BC mantiene valor predictivo fuerte, especialmente en segmentos de menores ingresos.

Figure 17: Gráfico de Prevalencia Multivariado

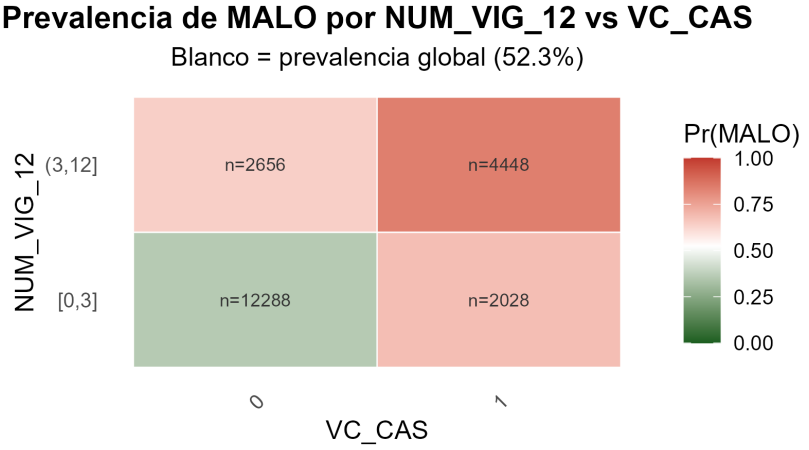
3.3.6 Prevalencia de Comportamiento Malo según Número de Meses Morosos y Deuda de Consumo



Se observa una prevalencia de comportamiento malo muy alta en todos los rangos de ambas variables. La única caracterización en donde se refleja un comportamiento bueno es cuando ambas variables tienen valores muy bajos (cercanos a 0).

Figure 18: Gráfico de Prevalencia Multivariado

3.3.7 Prevalencia de Comportamiento Malo según Número de Meses con Deuda Vigente e Indicador Meses con Deuda Vencida



Los clientes con más meses de deuda vigente (>3) y al menos un mes con deuda vencida o castigada (Deuda castigada = 1) presentan la mayor prevalencia de comportamiento malo (>70%). En cambio, quienes mantienen poca vigencia y sin castigos se concentran bajo la media. El historial de mora y vigencia prolongada son fuertes predictores de mal comportamiento.

Figure 19: Gráfico de Prevalencia Multivariado

3.4 Conclusiones Generales

Es importante tener en cuenta este análisis exploratorio para la creación de un modelo predictivo para la variable comportamiento.
Dentro de las relaciones encontradas, debemos tener en cuenta las variables que refieren al Indicador de meses con Deuda Vencida (VC CAS), Indicador de Boletín Comercial (IND BC), Indicador de IF (IND IF), Relación de Deuda Consumo (RPXDCON), y número de meses con deuda vigente y deuda morosa (NUM VIG 12 y NUM MOR 12). Estas variables son potenciales variables independientes predictoras, que podrían generar un modelo parsimonioso y descriptivo para la variable comportamiento. Hay que tener en cuenta la relación entre variables, para no generar multicolinealidad en las variables. Por lo tanto, se debe considerar si existen relaciones fuertes entre estas variables como ocurre en número de meses con deuda vigente y deuda morosa (NUM VIG 12 y NUM MOR 12), con una correlación fuerte mayor a 0.8.

4 Ajuste de modelo de clasificación para comportamiento crediticio

Al tener los registros limpios, completos y validos. Se requiere un modelo para clasificar de forma robusta el comportamiento económico de un cliente. Para esto, se utilizarán técnicas para seleccionar variables optimas como Forward, Stepwise y Backward. Sin embargo, paralelamente se realizará una selección de variables de manera manual, según las diferencias y asociaciones que presenten los comportamiento con las variables.

4.1 Modelo de regresión logística con selección de variables manual

4.1.1 Selección de variables

Ahora evaluaremos la relación entre las variables categóricas independientes y la variable dependiente Comportamiento. Aplicamos Chi-cuadrado de independencia, y cuando haya asociación significativa, reportamos V de Cramer.

Variable	Valor-p	V de Cramer
NUM_VIG_12	0	0.357
VC_CAS	0	0.353
NUM_MOR_12	0	0.349
IND_BC	0	0.298
IND_IF	0	0.245
NTCRED	0	0.153
NACR	0	0.119
VER_DOM	0	0.096

Table 17: Top 8 variables por V de Cramer (Chi-cuadrado frente a Comportamiento).

En este caso todas las variables presentan diferencias significativas al segmentarlas por Comportamiento, aunque es importante tener en cuenta que esta significación puede estar dada por la alta cantidad de observaciones. Para esto, solo incluiremos en el modelo aquellas presentan un V de cramer con una asociación ¿ 0.24. ("NUM_VIG_12", "NUM_MOR_12", "VC_CAS", "IND_BC", "IND_IF")

Para cuantitativas vs la variable de comportamiento usamos Kruskal–Wallis. Se mostrará el tamaño de efecto (epsilon-squared).

Variable	p-value	ε^2	Efecto
RPXDCON	0	0.173	grande ($\geq$0.14)
NUM_VIG_12	0	0.125	mediano ($\sim$0.06)
NUM_MOR_12	0	0.122	mediano ($\sim$0.06)
TMOT_06	3.82e$-$279	0.059	pequeño ($\sim$0.01)

Table 18: Top 4 por ε^2 (Kruskal–Wallis) frente a Comportamiento.

En este caso, ocurre lo mismo, en todas se encuentran diferencias significativas al 5%. Sin embargo, solo en una de ellas se presenta una asociación alta. En cambio, las demás de ellas mantienen un ϵ^2 realmente bajo, lo que nos sugiere que la fuerza de asociación entre las variables numéricas al separarlas por categorías de Comportamiento es baja al no haber suficiente variabilidad que las diferencie. Ahora, la variable con una diferencia significativa y alta está dada por "RPXDCON", la cual podría ser conveniente agregar al modelo.

A partir del análisis exploratorio se decidió excluir las variables "NUM VIG 12" y "NUM MOR12". Ambas presentan colinealidad entre sí y una correlación relevante con "RPXDCON" (relación de deuda de consumo). Dado que "RPXDCON" mostró diferencias claras de distribución según el comportamiento, su incorporación aporta mayor capacidad discriminante y evita redundancias, favoreciendo un modelo más parsimonioso y estable. Es por esto que las variables seleccionadas manualmente son: "VC_CAS", "IND_BC" y "RPDXCON", "IND_IF".

4.1.2 Ajuste del modelo y evaluación

Ajustando el modelo de clasificación por regresión logística, con variable dependiente "Comportamiento" y covariables "VC_CAS", "IND_BC" y "RPDXCON", "IND_IF". El comportamiento ajustado es el siguiente:

$$\log\left(\frac{P(\text{Malo})}{1-P(\text{Malo})}\right) = -1.1155 + 1.0177 \cdot \text{VC\_CAS\_1} + 1.2020 \cdot \text{IND\_BC\_1} + 1.3991 \cdot \text{RPXDCON} + 1.2330 \cdot \text{IND\_IF\_1}$$

4.1.3 Métricas del modelo

Las métricas del modelo especificado para comportamiento, mediante selección de variables manual son las siguientes:

Modelo	AUC	Accuracy	Sensitivity	Specificity	Precision	F1
modelo_seleccion	0.791	0.732	0.688	0.773	0.734	0.710

Table 19: Métricas del modelo seleccionado.

Las métricas indican el desempeño del modelo en términos de su capacidad para clasificar correctamente los casos de comportamiento crediticio. Donde, se pueden evaluar según métrica.

AUC (0.7913): El modelo presenta buena capacidad de discriminación. En términos probabilísticos, si tomamos

al azar un individuo Malo y uno Bueno, hay un 79.13% de probabilidad de que el modelo asigne un puntaje de riesgo más alto al Malo que al Bueno.
Accuracy (0.7324): El modelo clasifica correctamente el 73.24% del total de los casos, lo que indica un buen desempeño general.
Sensitivity (0.6879): El 68.79% de los individuos con comportamiento malo fueron correctamente identificados.
Specificity (0.7730): El 77.30% de los individuos con comportamiento bueno fueron correctamente clasificados como tal.
Precision (0.7341): De todas las predicciones que el modelo clasificó como comportamiento malo, el 73.41% fueron correctas.
F1 Score (0.7102): Es la media entre la precisión y el recall, indicando un buen equilibrio entre ambas.

4.2 Modelo de regresión logística con selección de variables mediante Forward, Backward y Stepwise

4.2.1 Selección de variables

La selección de variables se realiza mediante Forward, Stepwise y Backward. Estos métodos van agregando, eliminando o ambas (en el caso de stepwise) variables, buscando que estas variables minimicen el AIC. Al realizar esta selección de variables para el modelo los 3 métodos mencionados seleccionan las mismas variables, resultando en un mismo modelo final, por lo cual se utilizará solo uno.

4.2.2 Ajuste del modelo y evaluación

El modelo mantiene los siguientes coeficientes y variables.

Table 20: Coeficientes del modelo para comportamiento (Prob(Malo))

Coeficiente	Estimate	Valor-p	Coeficiente	Estimate	Valor-p
(Intercept)	-0.7943343	0.000	LINEACRNU	-0.0012706	0.000
SEXO1	-0.1130548	0.001	NUM_VIG_12	-0.0640592	0.000
EDAD	-0.0115336	0.000	NUM_MOR_12	0.1539192	0.000
EST_CIVILD	-0.0118620	0.906	VC_CAS1	0.4575140	0.000
EST_CIVILN	0.1773336	0.226	RPXLCNU	-0.4238123	0.000
EST_CIVILS	-0.1004271	0.009	RPXDCOM	1.3355475	0.000
EST_CIVILV	0.2005668	0.135	RPXDCON	1.9467813	0.000
NCC	0.0455664	0.033	IND_IF1	1.1018099	0.000
VER_TRA1	-2.0958942	0.000	IND_BC1	1.0720809	0.000
NDIR	0.0661683	0.000	TMOT_06	0.0158949	0.000
VER_DOM1	0.6463059	0.000	TDOC	0.0231674	0.000
NACR	0.0618432	0.117	NTCRED	0.2971147	0.000

Con la selección de variables mediante Forward, Stepwise y Backward, se seleccionan la mayoría de las variables, donde todas resultan ser significativamente distintas de 0 (con un nivel de significación del 0.05) excepto 4 de ellas: EST_CIVILD, EST_CIVILN, EST_CIVILV y NACR. Las 3 primeras correspondiendo a la misma variable categórica, donde, se sugiere que el estado civil no aporta significativamente a la clasificación del comportamiento, con solo una de sus categorías resultando significativa (soltero).

4.2.3 Métricas del modelo

El modelo ajustado para el comportamiento crediticio del cliente mediante la selección de variables por los metodos seleccionados, es el mismo, por lo que las métricas igual. Logrando las siguientes métricas.

Modelo	AUC	Accuracy	Sensitivity	Specificity	Precision	F1
modelo_backward	0.822	0.748	0.751	0.744	0.728	0.739

Table 21: Métricas del modelo *backward*.

Las métricas indican el desempeño del modelo en términos de su capacidad para clasificar correctamente los casos de comportamiento crediticio. Donde, se pueden evaluar según métrica.
AUC (0.8225): El modelo presenta buena capacidad de discriminación. En términos probabilísticos, si tomamos al azar un individuo Malo y uno Bueno, hay un 82.25% de probabilidad de que el modelo asigne un puntaje de riesgo más alto al Malo que al Bueno.
Accuracy (0.7476): El modelo clasifica correctamente el 74.76% del total de los casos, lo que indica un buen desempeño general.
Sensitivity (0.7514): El 75.14% de los individuos con comportamiento malo fueron correctamente identificados.
Specificity (0.7441): El 74.41% de los individuos con comportamiento bueno fueron correctamente clasificados como tal.
Precision (0.7279): De todas las predicciones que el modelo clasificó como comportamiento malo, el 72.79% fueron correctas.
F1 Score (0.7395): Es la media entre la precisión y el recall, indicando un buen equilibrio entre ambas.

4.3 Modelo final

Comparando las métricas de los modelos ajustados.

Modelo	AUC	Accuracy	Sensitivity	Specificity	Precision	F1
modelo_backward	0.822	0.748	0.751	0.744	0.728	0.739
modelo_seleccion	0.791	0.732	0.688	0.773	0.734	0.710

Table 22: Comparación de métricas entre modelos.

Podemos notar que en todos los casos, las métricas son mejores en el modelo por selección de variables por métodos de Forward, Stepwise y Backward. Excepto en Especificidad y Precisión, estos nos quiere decir que, la probabilidad de acertar una clasificación es mayor (precisión) en promedio para el comportamiento bueno y malo para el método de selección de variables manual. Comparando la matriz de confusión al entrenar con el 80% de los datos y 20% para validación de estos modelos.

	Real	
Predicción	**Bueno**	**Malo**
Bueno	1409	508
Malo	634	1734

Table 23: Matriz de confusión — *Modelo 1 (selección)*

	Real	
Predicción	**Bueno**	**Malo**
Bueno	1627	665
Malo	416	1577

Table 24: Matriz de confusión — *Modelo 2 (backward)*

Podemos notar que en la validación (20% para validación, 80% para entrenamiento) el modelo más simple (selección manual) entrega un mejor rendimiento para la precisión del comportamiento bueno de 73.5% contra 70.97% del modelo con selección por backward, stepwise y forward. Sin embargo, la precisión del modelo simple es menor para el comportamiento malo de 73.23% contra un 79.13% del modelo más complejo. Dado el mayor aumento y diferencia de 2.53% en la clasificación correcta del comportamiento bueno del modelo con selección de variables manual, y asumiendo una perdida de precisión del 5.9% para el comportamiento malo, se preferirá este modelo, a pesar de ser más optimista. Además, reforzando este modelo, es más interpretable para la clasificación de diferentes perfiles de clientes al tener solo 4 variables. Entonces, el modelo simple es algo más optimista que el modelo pesimista (mayor precisión en comportamiento malo) que es el modelo más complejo. Ahora bien, cada uno de los dos presenta características diferentes, con métricas optimas para la clasificación, por lo que el uso de cualquiera de los dos va en la estrategia de la entidad bancaria.

4.4 Evaluación de perfiles de clientes en el modelo para comportamiento crediticio

Como el modelo con métodos de selección mediante forward, backward y stepwise entregaron el mismo modelo, y además, contenía diversas variables e indices las cuales al definirlas, su interpretación es confusa y con baja comprensión. Se decide utilizar el modelo de selección de variables en la que se seleccionaron únicamente 4 variables ("RPXDCON"), ("VS_CAS"), ("IND_IF") y ("IND_BC"). Se definirán perfiles de clientes del banco a partir de estas 4 variables para ver el comportamiento, y como este es clasificado para el modelo con selección manual. Revisando también, la probabilidad de que este comportamiento sea Malo. (Decidiendo con umbral 0.5, si es mayor a este umbral el comportamiento es malo).

Tenemos el modelo:

$$\log\left(\frac{P(\text{Malo})}{1-P(\text{Malo})}\right) = -1.1155 + 1.018 \cdot \text{VC\_CAS\_1} + 1.202 \cdot \text{IND\_BC\_1} + 1.3961 \cdot \text{RPXDCON} + 1.233 \cdot \text{IND\_IF\_1}$$

- Perfil 1. Con deuda castigada, sin indicador BC, sin indicador IF y sin relación de deuda de consumo.

  $$VC\_CAS = \text{"1"}, IND\_BC = \text{"0"}, RPXDCON = 0, IND\_IF = \text{"0"}$$

  47.56% de probabilidad de mal comportamiento crediticio. (Buen comportamiento crediticio)

  Este perfil representa a un cliente con deuda castigada, pero sin indicadores BC y IF, ademas sin deuda de consumo. Estos factores no son suficientes para que el modelo le asigne un mal comportamiento, y la probabilidad de mal comportamiento es de 47.56%. Al estar por encima del umbral de 0.5, el sistema lo clasifica correctamente como "Bueno". Este tipo de cliente suele ser aprobado, pero podría ser sujeto a un monitoreo más estricto.

- Perfil 2. Sin deuda castigada, sin indicador BC, sin indicador IF y con relación de deuda de consumo de 0.5.

  $$VC\_CAS = \text{"0"}, IND\_BC = \text{"0"}, RPXDCON = 0.5, IND\_IF = \text{"0"}$$

  39.71% de probabilidad de mal comportamiento crediticio. (Buen comportamiento crediticio)

  Este es un cliente sin deudas castigadas, y sin indicadores BC e IF, y registra una relación de 0.5 de deuda de consumo maxima, es un cliente de riesgo moderado a bajo. Aunque tiene un historial de pago limpio (VC_CAS = "0"), la presencia de deuda de consumo eleva su perfil de riesgo. El modelo calcula su probabilidad de incumplimiento en 39.71%. Dado que este valor está por debajo del umbral de 0.5, el sistema lo clasifica como "Bueno".

- Perfil 3. Con deuda castigada, con indicador BC, con indicador IF y con relación de deuda de consumo maxima (1).

  $$VC\_CAS = \text{"1"}, IND\_BC = \text{"1"}, RPXDCON = 1, IND\_IF = \text{"1"}$$

  97.66% de probabilidad de mal comportamiento crediticio. (Mal comportamiento crediticio)

  Este es, inequívocamente, el perfil de mayor riesgo del grupo. Acumula todos los factores negativos: tiene un historial de deuda castigada (VC_CAS = "1") y también el indicador BC e IF (IND_BC = "1" y IND_IF = "1"), además de una relación de deuda de consumo máxima (1). El modelo refleja esto asignándole la probabilidad más alta de incumplimiento, un 97.66%. Es un claro comportamiento "Malo" y representa el tipo de cliente que el modelo busca identificar y rechazar.

- Perfil 4. Sin deuda castigada, sin indicador BC, sin indicador IF y sin relación de deuda de consumo.

  $$VC\_CAS = \text{"0"}, IND\_BC = \text{"0"}, RPXDCON = 0, IND\_IF = \text{"0"}$$

  24.68% de probabilidad de mal comportamiento crediticio. (Buen comportamiento crediticio)

  Este es el arquetipo del cliente "Bueno" y el perfil de más bajo riesgo. No tiene deudas de consumo vigentes, no tiene historial de deuda castigada, no tiene el indicador BC e IF. Como resultado, el modelo le asigna la probabilidad más baja de incumplimiento de todo el grupo, con solo un 24.68%. Su clasificación como "Bueno" es clara y representa un cliente seguro para la entidad.

- Perfil 5. Sin deuda castigada, sin indicador BC, sin indicador IF y con relación de deuda de consumo de 0.8.

  $$VC\_CAS = \text{"0"}, IND\_BC = \text{"0"}, RPXDCON = 0.8, IND\_IF = \text{"0"}$$

50.03% de probabilidad de mal comportamiento crediticio. (Mal comportamiento crediticio)

Este perfil representa a un cliente "en el límite" que no logra ser aprobado. Su principal característica es una alta carga de deuda de consumo (0.8), pero su factor más "sano" es un historial de pago limpio (VC_CAS = "0"), sin indicadores BC e IF. Debido a que tiene una alta carga crediticia de consumo, su probabilidad de incumplimiento se mantiene en 50.03%, por encima del umbral de 0.5, resultando en una clasificación de "Malo".

A modo general, la justificación de esas probabilidades se basa en el fuerte peso que el modelo le da a las variables VC_CAS, IND_BC, IND_IF y RPXDCON, en comparación con el riesgo base.
El modelo parte de un intercepto negativo (-1.1155), lo que significa que un cliente "perfecto" (con 0 en todas las variables, como el Perfil 2) tiene un riesgo base bajo.
El modelo identifica los cuatro factores de riesgo como altos: VC_CAS (1.018), IND_BC (1.202), RPXDCON (1.3961) e IND_IF (1.233). Los coeficientes son fuertemente positivos, aunque no tienen un peso tan grande para que cada uno por sí solo sea capaz de anular el intercepto negativo y empujar la probabilidad de clasificar un cliente como "Malo".
Teniendo lo anterior en cuenta, al no tener deuda vencida o castiga y no tener un indicador BC las probabilidades de clasificarse como un comportamiento "Malo" son menores, aun que, revisando el caso del perfil 5 (Sin castigo, Sin BC, sin IF, relación deuda de consumo 0.8) con una probabilidad de comportamiento "Malo" del 50.03, nos podemos dar cuenta de apesar de no tener estos castigos en la deuda y no tener indicador BC e IF, aun así, es posible clasificarse como un comportamiento "Malo", donde, si disminuimos esa relación de crédito de consumo, se clasificaría como "Bueno".

5 Conclusiones

Durante este trabajo se llegó a la disposición de dos modelos válidos, pero con distintos perfiles y estrategias de riesgo. El modelo backward (AUC = 0.825) es más conservador y pesimista: identifica más casos de mal comportamiento (Mayor sensibilidad), a costa de aumentar algo los falsos positivos. En cambio, el modelo de selección manual (AUC = 0.791) es más optimista: tiende a aprobar más (mayor especificidad), pero detecta menos comportamientos malos (menor sensibilidad). La elección de uno de estos modelos debe alinearse con la estrategia de riesgo, los costos por error y contexto de negocio.
Entonces, la forma en que se deben usar estos modelos depende de la situación. Donde, por ejemplo, si se está en un segmento o reunion de clientes aparentemente más riesgosos o nuevos, o bien, en contextos economicos adversos la politica debería ser proteger la cartera y usar el modelo conservador. Ahora bien, si el contexto económico es optimista, la estrategia busca nuevos clientes apesar del mayor riesgo y se tratan segmentos conocidos. El modelo a utilizar sería el modelo de selección manual (optimista). Aun que, lo recomendable seria maximizar las aprobaciones, y controlar ese riesgo moderado con algún método de garantía.
Es importante que el punto de corte (umbral) para el comportamiento malo es de 0.5, es decir, el umbral clasico y por defecto. Sin embargo, este umbral no es fijo, y es posible calibrarlo según los costos (manteniendo como objetivo minimizar los falsos positivos o negativos). Mover este umbral también permite que el modelo sea más conservador u optimista, sin reentrenarlo.
Respecto a los sesgos y limitaciones, tenemos que en un principio se eliminó GSE (por alta proporción de faltantes y falta de variables para clasificarlo), reduciendo el riesgo de sesgo. Algunas variables predictoras reflejan indicadores y practicas internas (ej. deuda vencida o castigada, indicador BC e IF) lo cual, es posible que existan errores al tabular si mantiene deuda castigada o un error al calcular estos indicadores.
En resumen, se mantienen estrategias duales: usar el modelo conservador donde importa protegerse (mayor detección de comportamiento malo) y el optimista donde interesa crecer con control (más aprobaciones con optima precisión). Complementando con un umbral dinámico, monitoreo y la próxima recalibración, la entidad bancaria obtendría una base adaptable que balancea el crecimiento y riesgo según contexto.

6 Librerías utilizadas

A continuación se detallan las principales librerías empleadas, las funciones clave y su propósito dentro del flujo de trabajo.

Paquete	Función	Utilización
knitr	`opts_chunk$set`	Configuración global de los *chunks* (ocultar mensajes y advertencias, opciones por defecto).
readxl	`read_excel`	Importación de datos desde archivos `.xlsx` al `data.frame` `scoring`.
dplyr	`select`	Selección/eliminación de columnas; creación de subconjuntos (p. ej., solo numéricas) y exclusión de `GSE` del análisis final.
	`mutate`	Creación y modificación de columnas; ingeniería de características (p. ej., grupo etario, estandarización).
	`filter`	Filtrado de filas según condiciones; eliminación de `NA` antes del modelado.
	`case_when`	Lógica condicional dentro de `mutate` (asignación de grupo etario a partir de `EDAD`).
	`arrange`	Ordenamiento de filas (p. ej., ranking de *outliers*).
	`summarise`	Cálculo de estadísticas resumen (media, desviación estándar) para estandarización y reportes.
	`count`	Conteo de frecuencias categóricas (gráficos de barras y tablas).
	`group_by`	Agrupación previa a operaciones como `summarise`.
	`across`	Aplicación de funciones a múltiples columnas (p. ej., `droplevels` sobre factores).
	`%>%`	Operador *pipe* para encadenar pasos de forma legible.
purrr	`imap_dfr`	Iteración sobre columnas y nombres a la vez; construcción de tablas resumen (p. ej., *outliers*).
	`map_dfr`	Aplicación de funciones a listas (pares de variables) y combinación en un único `data.frame`.
tidyr	`drop_na`	Eliminación de filas con `NA` en columnas específicas para gráficos/modelos sin errores.
plotly	`plot_ly`	Creación de gráficos interactivos (boxplots, barras, dispersión, pie).
	`layout`	Personalización de títulos, ejes y menús desplegables (`updatemenus`).
	`subplot`	Composición de múltiples gráficos en una sola visualización (p. ej., boxplot + densidad).
nnet	`multinom`	Ajuste de regresión logística multinomial (predicción de `GSE`).
caret	`createDataPartition`	Partición estratificada en entrenamiento/prueba.
	`trainControl`	Configuración del entrenamiento (p. ej., validación cruzada `cv`).
	`train`	Entrenamiento y ajuste de hiperparámetros (p. ej., `k` óptimo en k-NN).
	`confusionMatrix`	Evaluación de clasificación (Accuracy, Sensibilidad, Especificidad).
ggplot2	`ggplot`, `aes`	Inicio de gráficos y mapeo estético (ejes, color). Base de gráficos estáticos.
	`geom_point`, `geom_tile`, `geom_boxplot`	Dispersión, mapas de calor (correlación), boxplots.
	`facet_wrap`	Matrices de gráficos por subconjuntos.
	`labs`, `theme_minimal`, `theme`	Etiquetas y temas para legibilidad y estética.
	`scale_fill_gradient2`	Escala divergente para mapas de calor (-1 a 1).
	`coord_flip`	Boxplots horizontales para muchas categorías.
forcats	`fct_lump_n`	Agrupar niveles poco frecuentes en “Otros”.
	`fct_explicit_na`	Convertir `NA` en nivel explícito para visualización.
patchwork	`+`, `plot_layout`, `wrap_elements`	Composición de múltiples `ggplot2` y objetos (p. ej., tablas) en un diseño único.

Paquete	Función	Utilización
gridExtra	`tableGrob`, `ttheme_minimal`	Tablas como objetos gráficos integrables junto a `ggplot2`.
broom	`tidy`	Conversión de salidas de modelos (`glm`, etc.) a formato `tidy` para análisis/tablas.
pROC	`roc`, `auc`, `coords`	Curvas ROC, AUC y punto de corte óptimo.
gt	`gt`, `fmt_number`, `tab_header`	Tablas de presentación para resultados y perfiles de clientes.

Table 26: Librerías, funciones y uso dentro del flujo de análisis

7 Referencias

Parveen, K. R., Thangaraju, P. (2024). Enhanced credit scoring prediction using KNN-Z-Score based logistic regression (KZ-LR) algorithm. Journal of Electrical Systems, 20(3), 7230–7237.

AIM Chile. (2024). Actualización y Manual de Aplicación – GSE AIM 2023. Asociación de Investigadores de Mercado y Opinión Pública de Chile. https://aimchile.cl/wp-content/uploads/2025/06/Actualizacion-y-Manual-GSE-AIM-2023.pdf

GfK Chile. (2019). Los nuevos GSE [Documento informativo]. GfK Chile. https://cdn2.hubspot.net/hubfs/2405078/cms-pdfs/fileadmin/user$_u pload/country_o ne_p ager/cl/g f k_g se_1 90502_f inal.pdf$

Rodríguez, P., Valenzuela, J. P., Truffello, R., Ulloa, J., Matas, M., Quintana, D., Hernández, C., Muñoz, C., Requena, B. (2019). Un modelo de identificación de requerimientos de nueva infraestructura pública en educación básica (Informe final FONIDE). Centro de Estudios, Ministerio de Educación (Chile). https://centroestudios.mineduc.cl/wp-content/uploads/sites/100/2021/08/061-Rodriguez-FINAL.pdf

Agresti, A. (2019). An Introduction to Categorical Data Analysis (3rd ed.). Wiley.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates.

Conover, W. J. (1999). Practical Nonparametric Statistics (3rd ed.). Wiley.

Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

Field, A. (2013). Discovering Statistics Using IBM SPSS Statistics (4th ed.). SAGE Publications.

Olejnik, S., Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. Psychological Methods, 8(4), 434–447.