Análisis exploratorio y ajuste de modelo para comportamiento crediticio de los clientes

Base de datos Scoring

Felipe Neira Rojas & Angel Llanos Herrera

Profesor: José Zúñiga Núñez.
Econometría [IES-424]

Universidad Católica del Maule

23 de octubre, 2025

Índice

Contexto

- **Definición:** El scoring en econometría es un conjunto de métodos estadísticos/modelos que asigna a cada solicitante o cliente un puntaje de riesgo.
- **Salida del modelo:** Ese puntaje resume la probabilidad de observar un comportamiento BUENO o MALO en una ventana futura.
- **Objetivo práctico:** Permitir decisiones consistentes y escalables en gestión crediticia. Traduciendo datos históricos del cliente en una medida cuantitativa y comparable de riesgo esperado para apoyar políticas y procesos operativos
- **Decisiones:** Aprobar o Rechazar un crédito: fijar límites y tasas de interés, priorizar campañas comerciales, gestionar cobranzas.

Datos

- **Datos de partida:** Se utilizan datos históricos (sociodemográficos, bancarios, renta, deudas, etc.) más una etiqueta de desempeño (ej., BUENO/MALO).
- **Modelo de clasificación:** Con esas bases se entrena un modelo que estima la probabilidad de un comportamiento (BUENO/MALO) dado un conjunto de variables predictoras.
- **Base empleada:** Se dispone de una base de datos scoring con 21.520 observaciones de clientes y 29 variables. (se eliminan al azar 100 obs)

Variable	Descripción
VC_CAS	Indicador de 1+ meses (últimos 12) con deuda vencida o castigada.
IND_BC	Indicador del BC en los últimos 6 meses.
IND_IF	Indicador del IF en los últimos 6 meses.
RPXDCON	Relación de deuda de consumo (promedio a máximo).
GSE	Nivel socioeconómico del cliente.
RENTA	Renta mensual promedio del cliente.
AVALUO	Avalúo de bienes raíces.

Cuadro 1 – Variables seleccionadas del diccionario.

Valores Faltantes

- **Depuración inicial de categorías y faltantes:** Se partió detectando registros no nulos en todas las variables, pero se identificó la necesidad de capturar faltantes por errores de tabulación (ej. “NA”, “nA”, “na”, etc.). Por ello se revisaron las categorías de variables categóricas.

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 %

Cuadro 2 – Distribución de *GSE*: conteo y porcentaje.

- Estándar AIM Chile (2024): GSE se calcula con un índice socioeconómico del hogar a partir de (i) ingreso per cápita equivalente (ingreso ajustado por tamaño del hogar con escala $n^{0,7}$), (ii) nivel educativo del principal sostenedor y (iii) ocupación del principal sostenedor; el modelo se calibra con datos de CASEN y EPF.

Valores atípicos

Valores atípicos (criterio de Tukey): Cuantificarlos univariadamente es clave para detectar variables que pueden mantener distorsión en media y varianza, sesgar las relaciones del modelo y dificultar la detección de patrones.

Valores atípicos

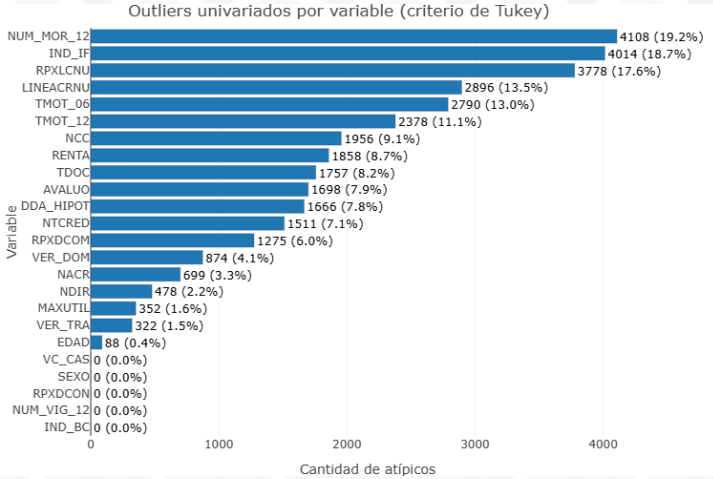

Figura 1 – Ranking de variables con valores atípicos.

Tratamiento de valores faltantes en grupo socioeconómico (GSE)

- **Disponibilidad de predictores para GSE:** La base no contiene las variables que definen GSE, RENTA es la mejor variable disponible. Igual se explorarán diferencias y asociaciones con otras variables para no perder señal útil.
- **Estrategia de imputación (validación previa):** Se entrenará una regresión multilogística para predecir GSE, aunque solo se usará para imputar si logra $>70\,\%$ de precisión por categoría en validación (criterio mínimo).
- **Método preferente de imputación:** La evidencia sugiere que KNN preserva mejor las relaciones entre variables y mejora el rendimiento frente a imputaciones por media/mediana/moda en contextos de credit scoring (Parveen & Thangaraju, 2024).

Selección de Variables para Tratamiento

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

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

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

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

Selección de Variables para Tratamiento

- **Asociaciones fuerte:** Entre las numéricas, RENTA y AVALUO muestran asociación fuerte y consistente con GSE.
- **Redundancia/colinealidad:** Las dos variables mencionadas están altamente correlacionadas, usarlas juntas introduciría posible multicolinealidad, inestabilizando los pesos, logrando degradar el rendimiento.
- Se prioriza RENTA y se excluye AVALUO, favoreciendo un modelo más parsimonioso y robusto.

Validación de Imputación de Datos Faltantes

- Validación del modelo de regresión multilogística: Se ajustó una regresión multilogística para GSE (excluyendo NA) con RENTA como predictor, usando un división 80/20 (entrenamiento/validación).

	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 %

Cuadro 5 – Matriz de precisión por clase predicha regresión multilogística..

- Rendimiento pobre e inutilizable de manera válida ($< 70\,\%$ de precisión por clase).

Imputación por KNN

- KNN para GSE (validación 80/20): Se clasificó GSE con RENTA como única predictora usando KNN y seleccionando k por máximo accuracy; el óptimo fue k=31 con accuracy = 0.477 (desempeño pobre), y se reporta la matriz de confusión del 20 % de validación.

	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 %

Cuadro 6 – Matriz de precisión por clase predicha (porcentaje por fila)..

- Rendimiento y precisión aumenta con KNN contra regresión multilogística, pero aún insuficiente para imputar de manera válida ($< 70\,\%$ de precisión por clase).

Decisión sobre Imputación de Datos

- **Imputación y sesgo:** Cuando el faltante es pequeño, imputar suele conservar distribuciones y relaciones. Aun que, con porcentajes altos y, si los faltantes son "completamente al azar", la imputación introduce sesgos (empuja estimaciones hacia valores demasiado optimistas o conservadores).
- **Regla práctica clave:** Antes de imputar, bajar el faltante total a aproximadamente 10 % o menos, eliminando variables que concentran la mayor parte de los vacíos siempre que no comprometa la validez del estudio (Romero-Duque et al., 2023).
- **Aplicación al caso (GSE):** Con $>50\,\%$ de faltantes y modelos que no logran una clasificación/imputación robusta, usar GSE no es válido: implicaría alto sesgo. Por tanto, se excluye GSE del análisis.

Análisis Exploratorio Univariado

Comportamiento	Porcentaje
Malo	52.33 %
Bueno	47.67 %

Cuadro 7 – Frecuencia Relativa Porcentual del Comportamiento del Cliente.

Variable de Interés presenta balanceo en sus comportamiento, en donde el comportamiento malo tiene una mayor dominancia con un 52,33 % de clientes.

Análisis Exploratorio Univariado

Sexo	Porcentaje
Masculino	51.53 %
Femenino	48.47 %

Cuadro 8 – Frecuencia relativa del sexo del cliente.

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

Cuadro 9 – Distribución porcentual del estado civil.

Análisis Exploratorio Univariado

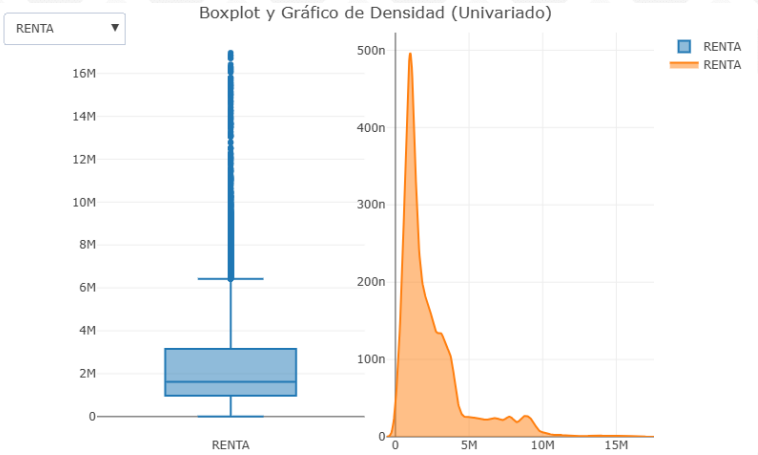

Figura 2 – Gráfico distribucional de la renta de los clientes.

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

Cuadro 10 – Resumen descriptivo de la variable renta.

Análisis Exploratorio Univariado

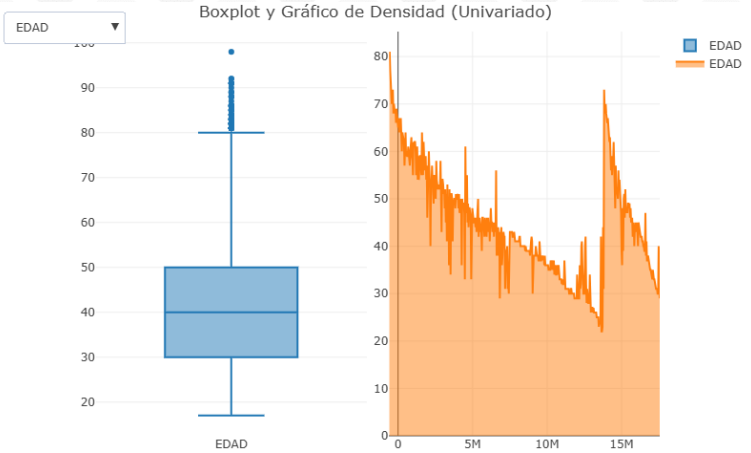

Figura 3 – Gráfico distribucional de la edad de los clientes.

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

Cuadro 11 – Resumen descriptivo de la variable edad.

Análisis Exploratorio Bivariado

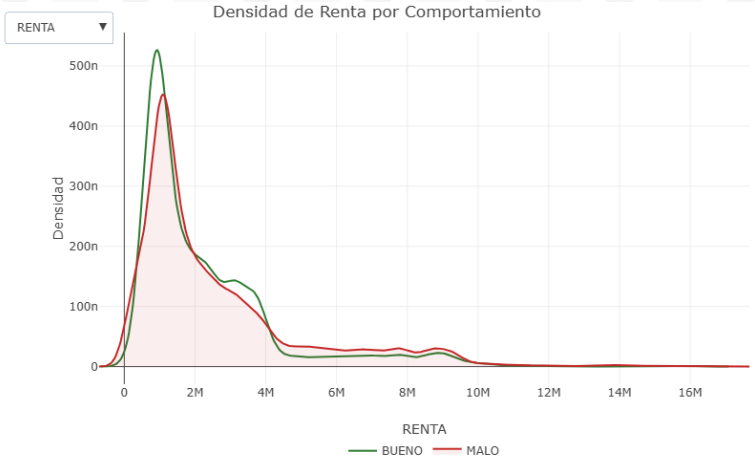

Figura 4 – Gráfico distribucional de renta por comportamiento.

Análisis Exploratorio Bivariado

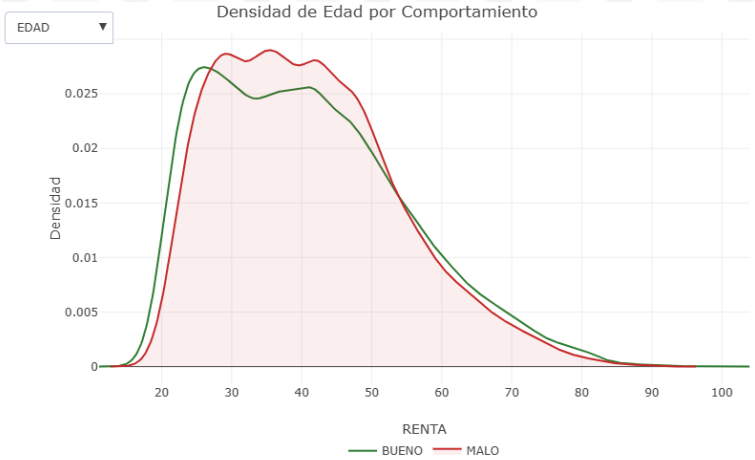

Figura 5 – Gráfico distribucional de edad por comportamiento.

Análisis Exploratorio Bivariado

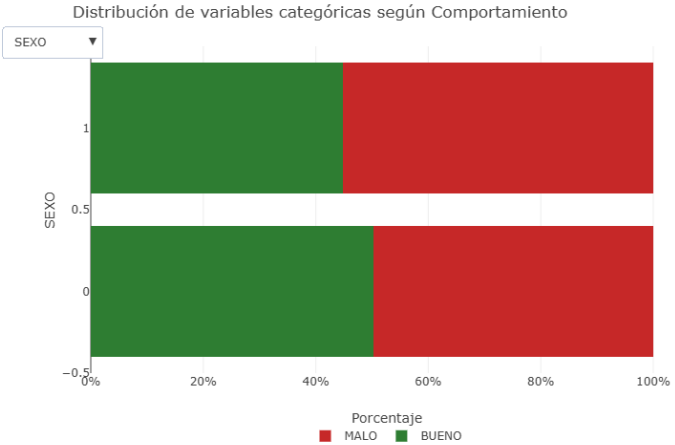

Figura 6 – Gráfico distribucional de sexo por comportamiento.

Análisis Exploratorio Bivariado

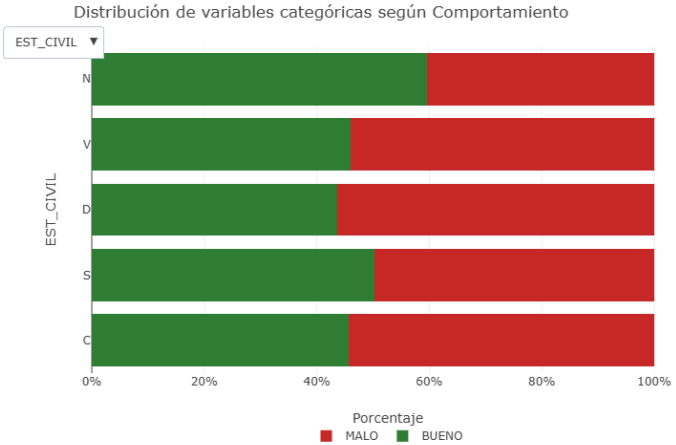

Figura 7 – Gráfico distribucional de estado civil por comportamiento.

Análisis Exploratorio Bivariado

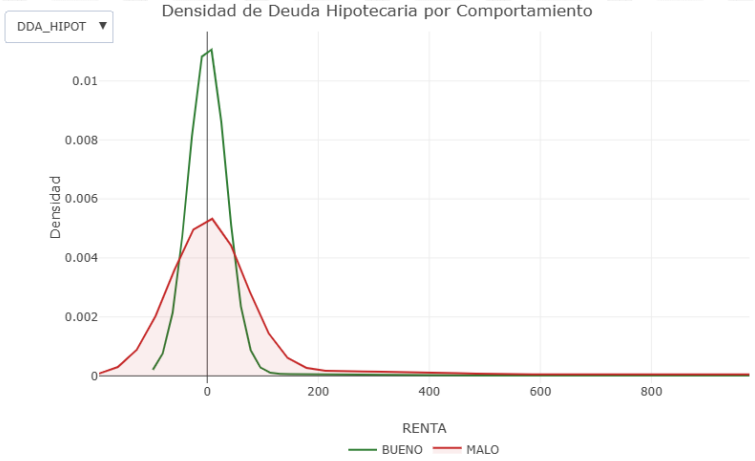

Figura 8 – Gráfico distribucional de deuda consumo por comportamiento.

Análisis Exploratorio Multivariado: Relaciones entre Variables

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

Cuadro 12 – Principales asociaciones entre variables según coeficiente de relación.

Análisis Exploratorio Multivariado: Relaciones con Comportamiento

Variable	Asociación
Indicador de 1+ meses con deuda vencida o castigada (últimos 12 m)	0.353
Indicador en boletín comercial (últimos 6 m)	0.298
Indicador en instituciones financieras (últimos 6 m)	0.245
Relación de deuda de consumo promedio a máximo	0.171
Meses con deuda vigente (últimos 12 m)	0.118
Meses con deuda morosa (últimos 12 m)	0.101

Cuadro 13 – Asociación de cada variable con comportamiento.

Análisis Exploratorio Multivariado

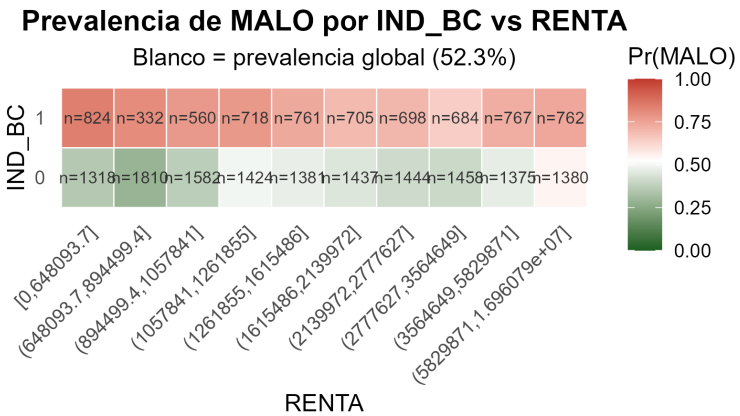

Figura 9 – Gráfico de Prevalencia Multivariado.

Análisis Exploratorio Multivariado

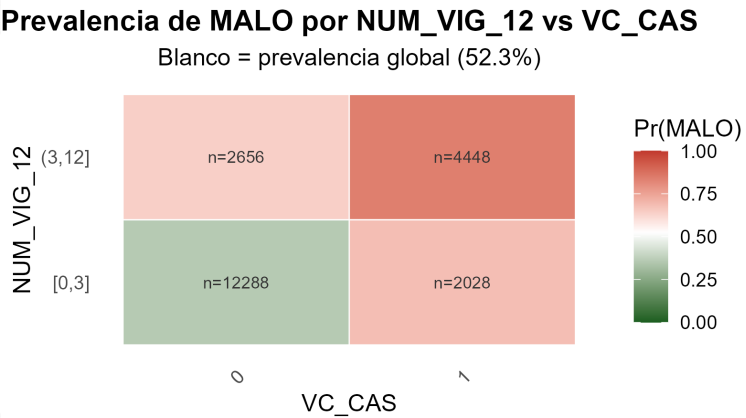

Figura 10 – Gráfico de Prevalencia Multivariado.

Selección de Variables

- Evaluaremos la relación entre las variables categóricas independientes y la variable dependiente COMPORTAMIENTO.

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

Cuadro 14 – Top 8 variables por V de Cramer (Chi-cuadrado frente a *COMPORTAMIENTO*)..

- Solo se incorporan las con V de Cramer > 0.24 para asegurar relevancia práctica: NUM_VIG_12, NUM_MOR_12, VC_CAS, IND_BC e IND_IF.

Criterios de Selección

- 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)

Cuadro 15 – Top 4 por ε^2 (Kruskal–Wallis) frente a *COMPORTAMIENTO.*.

- Aunque todas son significativas al 5 %, solo RPXDCON muestra asociación alta. Se incluye RPXDCON en el modelo y se descartan las demás numéricas por baja diferenciación.

Decisión

- **Variables seleccionadas en principio:** NUM_VIG_12, NUM_MOR_12, VC_CAS, IND_BC, IND_IF y RPXDCON.
- **Variables finales seleccionadas:** VC_CAS, IND_BC, IND_IF y RPXDCON.
- La selección de estas 6 variables en principio concuerdan con las que tienen mayor asociación con el comportamiento del cliente. La elección final se define por la gran relación entre NUM_VIG_12 y NUM_MOR_12, en conjunto también con RPXDCON, donde esta última tiene mayor poder discriminatorio, según el análisis exploratorio

Resultado del Modelo y Métricas

- Ajustando el modelo de clasificación por regresión logística, con variable dependiente ÇOMPORTAMIENTOż covariables "VC_CAS", ÏND_BCż RPDXCON", ÏND_IF".

$$\log\left(\frac{P(\text{MALO})}{1-P(\text{MALO})}\right) = -1{,}11 + 1{,}02 \cdot \text{VC\_CAS\_1} + 1{,}2 \cdot \text{IND\_BC\_1} + 1{,}40 \cdot \text{RPXDCON} + 1{,}23 \cdot \text{IND\_IF\_1}$$

- Las métricas son las siguientes:

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

Cuadro 16 – Métricas del modelo seleccionado..

Modelo con selección de variables mediante Forward, Backward y Stepwise

- **Selección por AIC (Forward/Stepwise/Backward):** se aplicaron los tres métodos y todos convergieron al mismo conjunto de variables, minimizando el AIC y produciendo un único modelo final.

Cuadro 17 – 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

Métricas del Modelo

- Las métricas son las siguientes:

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

Cuadro 18 – Métricas del modelo *backward*..

Comparación del Modelo con Selección Manual y Automática

- 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

Cuadro 19 – Comparación de métricas entre modelos..

- Los modelos con selección automática (Forward/Stepwise/Backward) superan en la mayoría de las métricas al modelo por selección manual. Excepto en la especificidad y la precisión, que son son más altas en el modelo manual.

Comparación del Modelo con Selección Manual y Automática

- 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

Cuadro 20 – Matriz de confusión — *Modelo 1 (selección)*.

	Real	
Predicción	**BUENO**	**MALO**
BUENO	1627	665
MALO	416	1577

Cuadro 21 – Matriz de confusión — *Modelo 2 (backward)*.

- **Precisión por clase (validación 20 %)**:
  BUENO 73.5 % (simple) vs 71.0 % (complejo)
  MALO 73.23 % (simple) vs 79.13 % (complejo).

Perfiles de Cliente

- Utilizando el modelo simple (selección de variables manual), se crean diferentes perfiles para poner aprueba el modelo.

$$\log\left(\frac{P(\text{MALO})}{1-P(\text{MALO})}\right) = -1{,}11 + 1{,}02 \cdot \text{VC\_CAS\_1} + 1{,}20 \cdot \text{IND\_BC\_1} + 1{,}40 \cdot \text{RPXDCON} + 1{,}23 \cdot \text{IND\_IF\_1}$$

- Indicador del BC en los últimos 6 meses.
- Indicador de 1 o más meses con Deuda Vencida.
- Relación de Deuda Consumo.
- Indicador de IF en los últimos 6 meses.

Perfiles de clientes

Perfil	Prob(MALO)	Comportamiento
Perfil 1. Con deuda castigada, sin indicador BC, sin indicador IF y sin relación de deuda de consumo.	47.56 %	BUENO
Perfil 2. Sin deuda castigada, sin indicador BC, sin indicador IF y con relación de deuda de consumo de 0.5.	39.71 %	BUENO
Perfil 3. Con deuda castigada, con indicador BC, con indicador IF y con relación de deuda de consumo máxima (1).	97.66 %	MALO
Perfil 4. Sin deuda castigada, sin indicador BC, sin indicador IF y sin relación de deuda de consumo.	24.68 %	BUENO
Perfil 5. Sin deuda castigada, sin indicador BC, sin indicador IF y con relación de deuda de consumo de 0.8.	50.03 %	MALO

Cuadro 22 – Perfiles y probabilidad de mal comportamiento crediticio..

Perfiles de Cliente

- **Peso de predictores clave:** Las probabilidades se explican por el fuerte efecto de VC_CAS, IND_BC, IND_IF y RPXDCON frente al riesgo base.
- **Riesgo base bajo:** El intercepto negativo (-1.1155) implica que un cliente "perfecto" (ej. con 0 en todas las variables, como el Perfil 2) parte con probabilidad baja de mal comportamiento crediticio.
- **Efectos positivos pero acumulativos:** Los coeficientes positivos aumentan la probabilidad de mal comportamiento, aunque por sí solos no siempre superan el efecto del intercepto, pero en conjunto sí pueden empujar la probabilidad hacia el mal comportamiento.
- **Ausencia de variables que disminuyan probabilidad de mal comportamiento:** Se identifican como principales covariables factores de riesgo, pero no factores mitigadores.

Conclusiones

- **Dos modelos, dos estrategias:**
  - Conservador (backward, AUC=0.825): mayor sensibilidad (detecta más "MALO").
  - Optimista (selección manual, AUC=0.791): mayor especificidad (aprueba más), menor detección de "MALO".
- **Cuándo usar cada uno:**
  - Proteger cartera: segmentos nuevos/riesgosos o ciclo adverso: modelo conservador.
  - Crecer con control: ciclo favorable/segmentos conocidos: modelo optimista (límites, garantías, tasas de interés por riesgo).
- **Umbral ajustable (0.5 no es fijo):**
  Ajustar el punto de corte según costos de error (FP/FN) permite volver el mismo modelo más conservador u optimista sin reentrenar.

Conclusiones

- **Sesgos y limitaciones adicionales:**
  Se excluye GSE por faltantes altos y falta de predictores; algunos predictores reflejan prácticas internas (ej. castigos, indicadores BC/IF) susceptibles a errores de tabulación o cálculo.
- **Recomendación operativa (estrategia dual):**
  Implementar modelo conservador donde prima la contención de pérdidas y modelo optimista donde prima el crecimiento.

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.
- GfK Chile. (2019). Los nuevos GSE [Documento informativo]. GfK Chile.
- 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).

Referencias

- 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.