LICENCIATURA EN ESTADÍSTICA

Predicción de fraude financiero

Un análisis mediante modelos lineales generalizados

Autores: Franco Santini - Nicolas Gamboa - Andrés Roncaglia

Docentes: Boggio Gabriela - Harvey Guillermina - Costa Victorio

2024

Tabla de contenidos

Introducción **1**

Variables: **1**

Análisis descriptivo **2**

Modelado **8**

Modelo estimado 11

Analisis de residuos 12

Evaluación de la componente sistemática 12

Comprobación de la distribución propuesta 13

Interpretaciones **13**

Capacidad predictiva **15**

Discusión **17**

Introducción

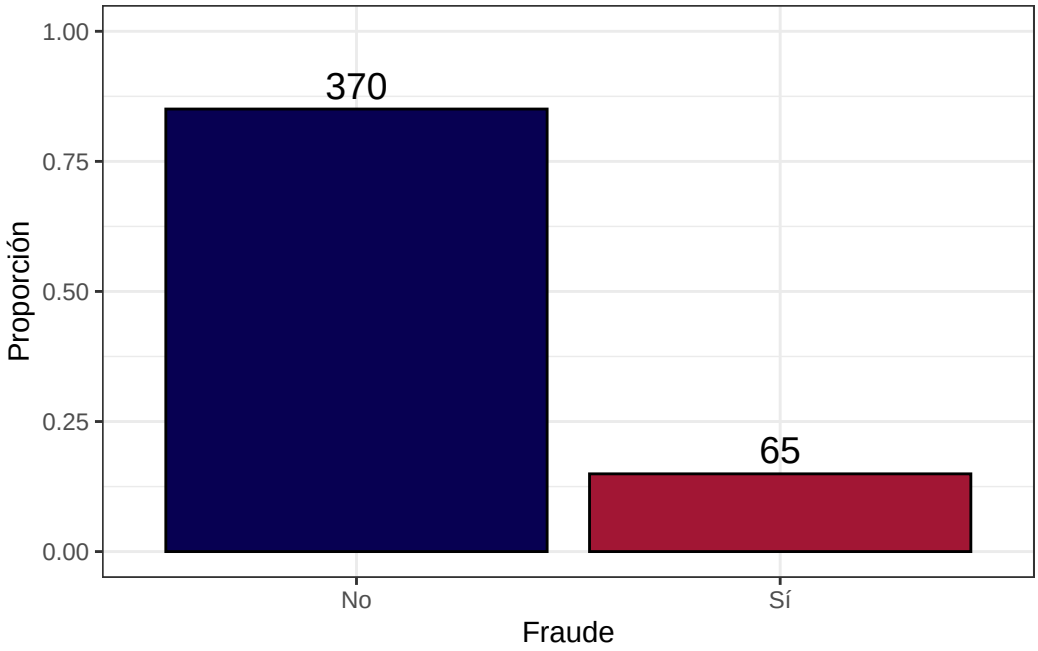
El fraude con tarjetas de crédito es una de las principales amenazas que sufren los bancos. Con el auge las transacciones digitales y los medios de pago electrónicos, se facilitaron los traspasos de dinero, pero junto con las ventajas también vinieron las consecuencias, y es que los métodos de fraude se han vuelto más sofisticados, generando pérdidas significativas a los bancos y afectando la confianza de los usuarios. Actividades como el uso no autorizado de tarjetas, la clonación de datos y transacciones fraudulentas requieren el desarrollo de tecnologías avanzadas para la detección temprana y la prevención.

Variables:

- `fraud_bool`: Indicadora de si la transacción fue fraude o no
- `income`: Ingreso anual en cuantiles
- `name_email_similarity`: Similitud del nombre en el email y el nombre del solicitante
- `customer_age`: Edad del cliente en décadas
- `days_since_request`: Días desde la solicitud
- `payment_type`: Tipo del plan de pago
- `employment_status`: Estado de empleo del solicitante
- `credit_risk_score`: Score de riesgo de la aplicación
- `email_is_free`: Tipo del dominio del email del aplicante (email pago o gratis)
- `housing_status`: Estado residencial del aplicante
- `phone_home_valid`: Validez del telefono fijo provisto
- `phone_mobile_valid`: Validez del telefono movil provisto
- `has_other_cards`: Indicadora de si la persona tiene otra tarjeta en el mismo banco
- `proposed_credit_limit`: Crédito limite propuesto por el aplicante
- `foreign_request`: Indicadora de si la solicitud fue hecha en el pais del banco
- `device_os`: Sistema operativo del dispositivo desde el que se hizo la solicitud
- `keep_alive_session`: Indicadora de si el solicitante decidió mantener la sesión iniciada al ingresar
- `device_distinct_emails_8w`: Número de emails distintos en la página del banco desde el mismo dispositivo usado en las últimas 8 semanas

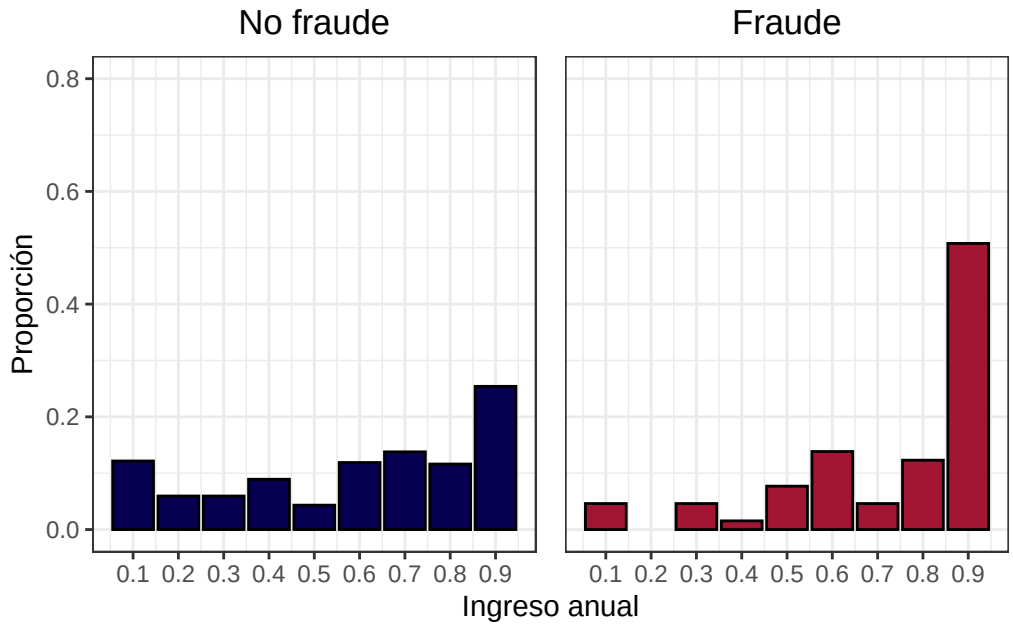
Análisis descriptivo

Figura 1: Proporción de clientes que comentieron fraude



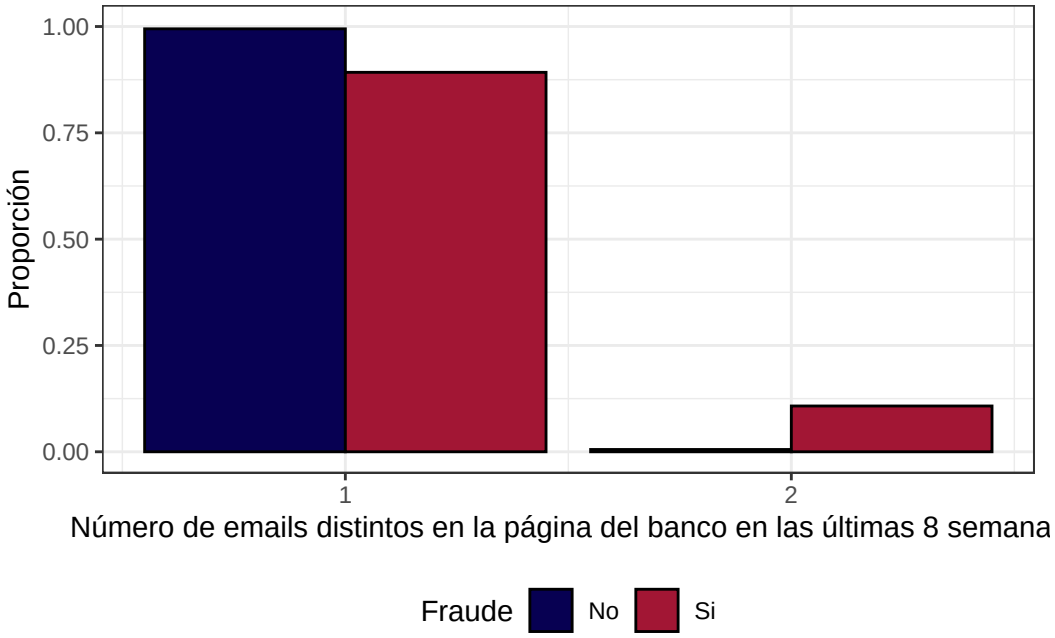
Es importante notar que las clases para la variable de interés son bastante desbalanceadas, siendo los clientes que cometen fraude poco frecuentes.

Figura 2: Distribución del ingreso anual según si la transacción es fraudulenta o no



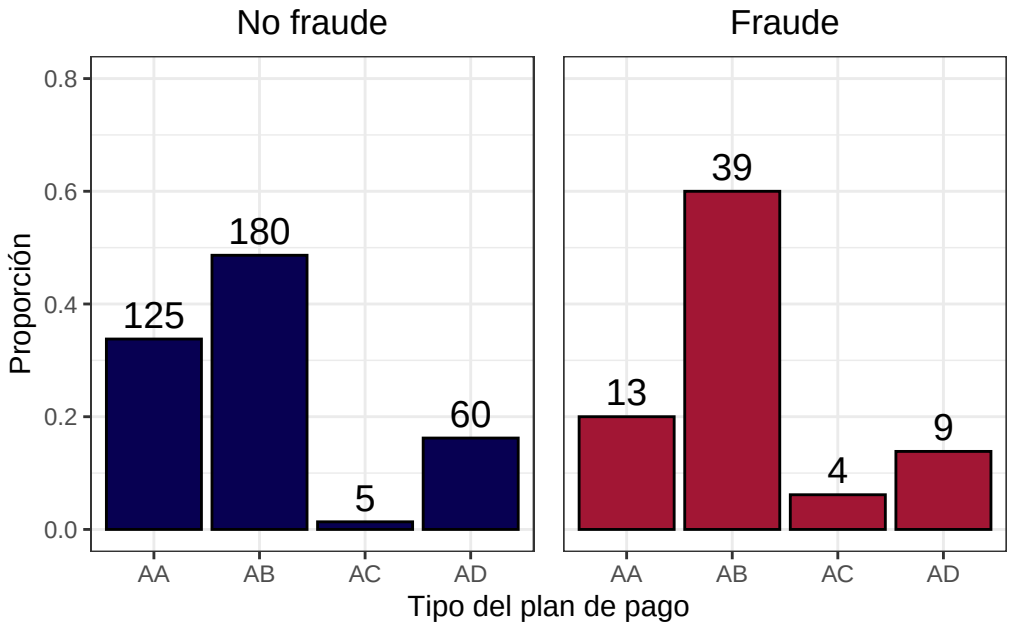
Se puede observar que las personas que cometieron fraude tienden a tener un ingreso anual registrado mayor. La distribución tiene una mayor asimetría a la izquierda.

Figura 3: Número de e-mails distintos en la página del banco del cliente en las últimas 8 semanas según si la transacción es fraudulenta o no



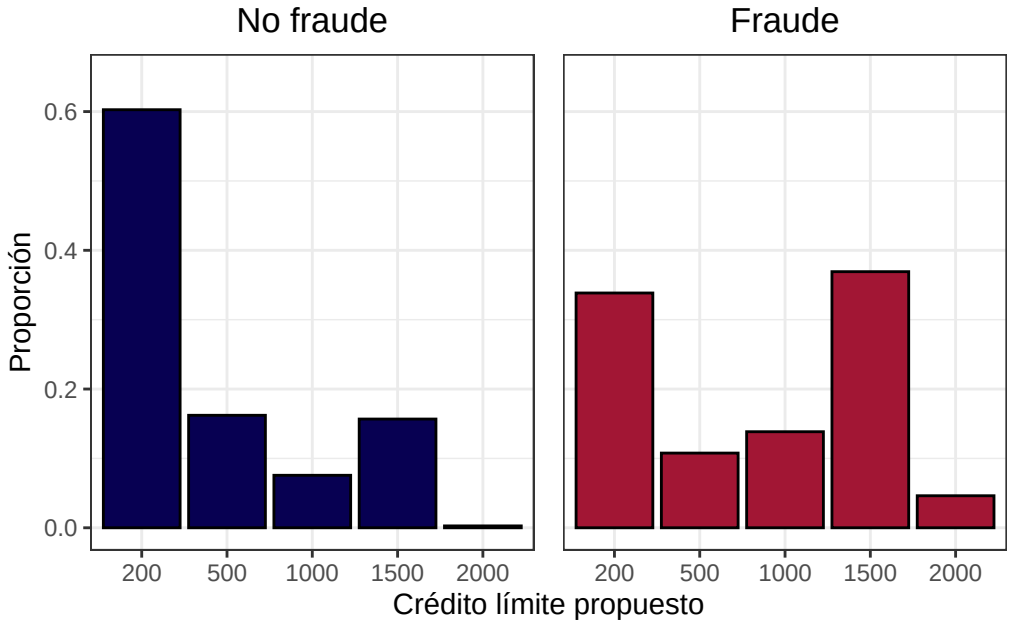
Parece ser que los clientes que cometen fraude son más propensos a tener 2 e-mails registrados en la página del banco que los que no cometen fraude.

Figura 4: Proporción del tipo de pago según si la transacción es fraudulenta o no



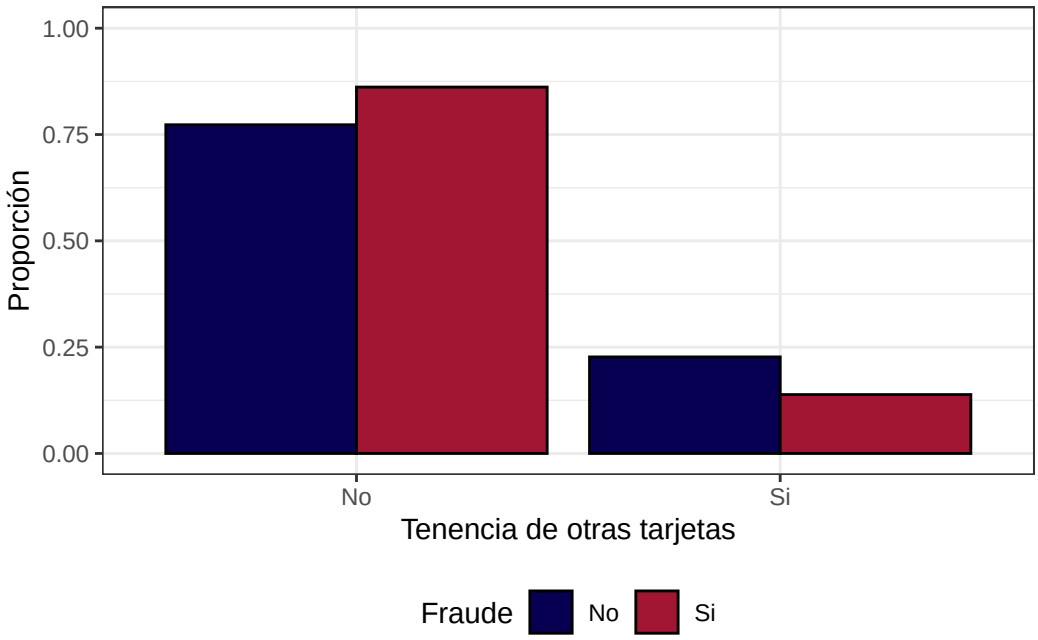
En general, las personas que cometen fraude parecen preferir los métodos de pago "AB" y "AC" por encima del resto, al contrario de las personas que operan de manera legítima que prefieren de igual manera los tipos de pago "AA", "AB" y "AC". Se puede notar también que la forma de pago "AE" no es muy popular.

Figura 5: Distribución del límite crediticio propuesto por el solicitante según si la transacción es fraudulenta o no



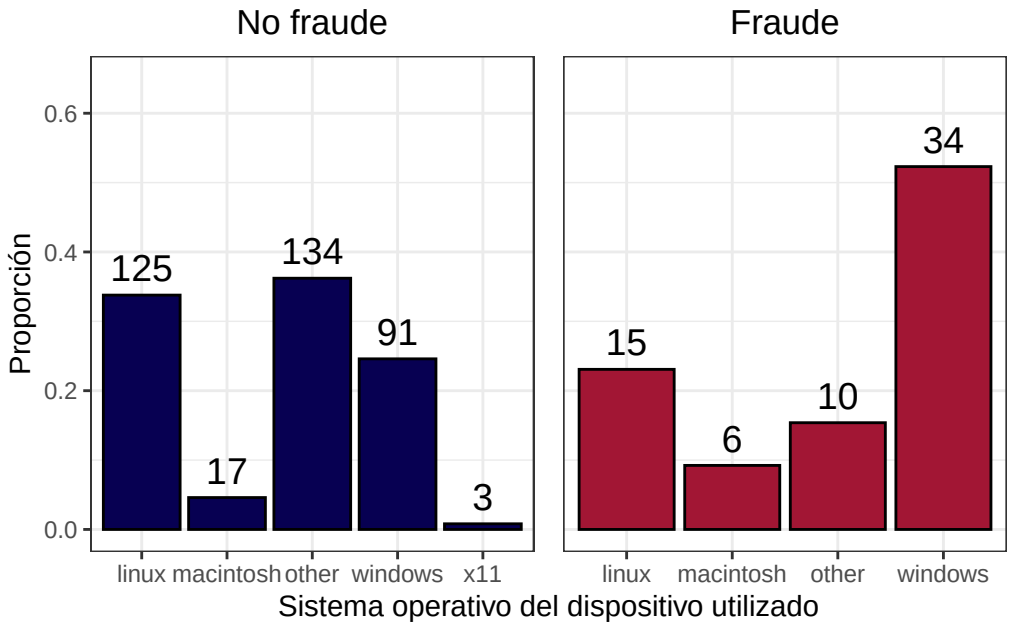
Se puede destacar en este gráfico que las personas que cometen fraude son ligeramente más propensas a pedir créditos más altos.

Figura 6: Proporción de la tenencia de otra tarjeta en el mismo banco según si la transacción es fraudulenta o no



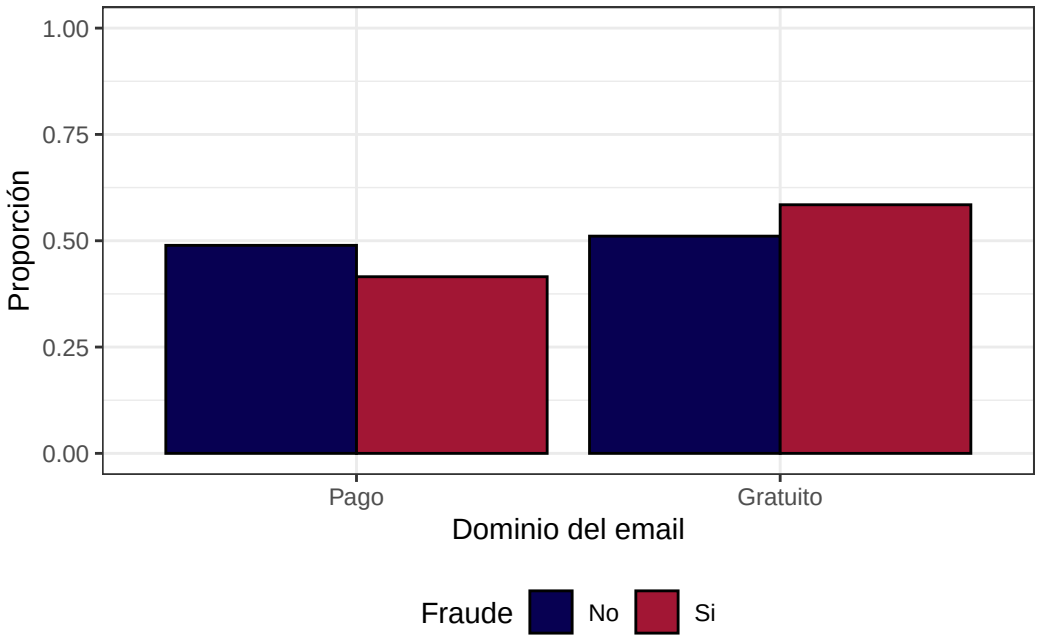
En cuanto a la tenencia de otra tarjeta en el mismo banco, suele ser común no poseer otra, sin embargo las personas que operan de forma legal se inclinan a tener más de una tarjeta en mayor medida que aquellos que cometen fraude.

Figura 7: Proporción del sistema operativo del dispositivo según si la transacción es fraudulenta o no



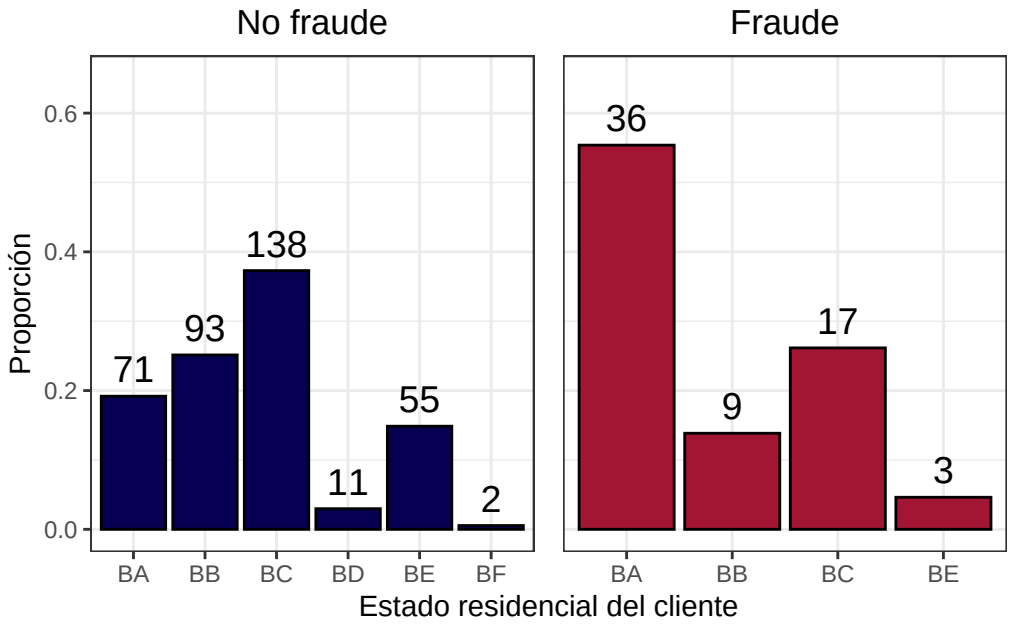
Se puede observar que las personas que cometen fraude, parecen utilizar Windows en mayor proporción que las personas que no cometen fraude. Dado que muy pocos clientes utilizan el sistema operativo "x11" se decidió agregarlo a la categoría "other".

Figura 8: Proporción del tipo de dominio del email según si la transacción es fraudulenta o no



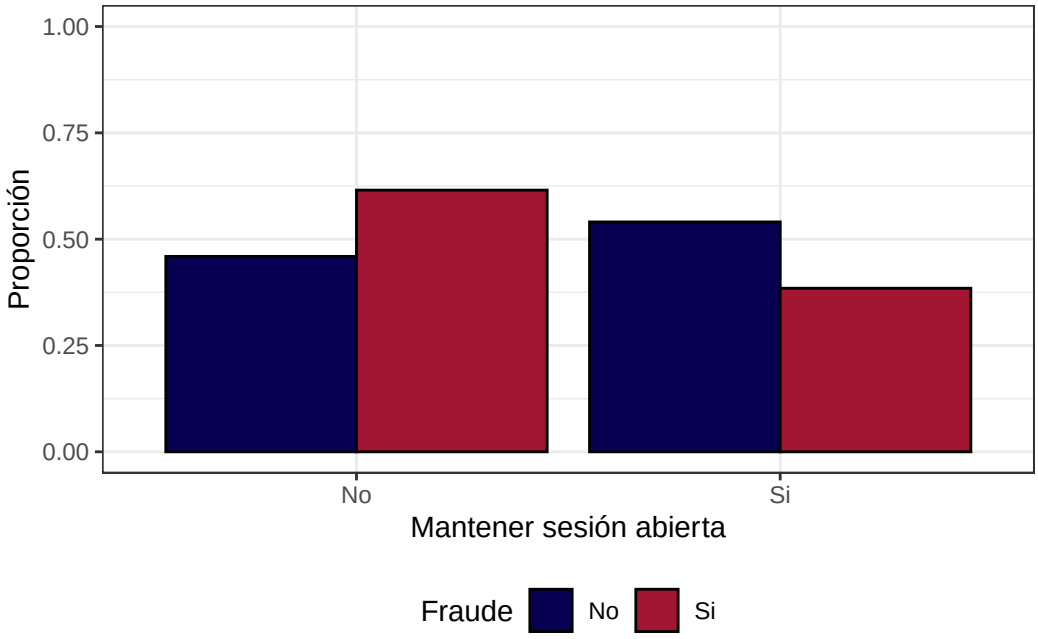
También se puede destacar que las operaciones fraudulentas parecen ser más comunes cuando el dominio del email del solicitante es gratuito que cuando es pago.

Figura 9: Estado residencial del cliente según si la transacción es fraudulenta o no



Se puede notar que los clientes que cometen fraude suelen tener un estado residencial "BA" en mayor proporción que aquellos clientes que no cometieron fraude, ya que estos últimos suelen tener un estado residencial "BC" y "BB" en mayor medida que "BA". Dado que hay pocos datos en las categorías "BD", "BE" y "BF" se decidió agruparlas en una sola categoría llamada "Otros".

Figura 10: Proporción de opciones de inicio de sesión según si la transacción es fraudulenta o no



Por lo general, cuando las transacciones son fraudulentas la persona decide no mantener la sesión abierta en la cuenta del banco en mayor proporción que cuando las transacciones son legítimas.

Modelado

Teniendo todo esto en cuenta se buscó un modelo que ajuste bien a los datos, para esto primero se realizó una selección de variables paso a paso, obteniendo el siguiente modelo:

$$logit(\pi_i) = \beta_0 + \beta_I \cdot I_i + \beta_{H1} \cdot H_{1i} + \beta_{H2} \cdot H_{2i} + \beta_{H3} \cdot H_{3i} + \beta_{Ph} \cdot Ph_i +$$
$$+\beta_{Pm} \cdot Pm_i + \beta_C \cdot C_i + \beta_L \cdot L_i + \beta_{D1} \cdot D_{1i} + \beta_{D2} \cdot D_{2i} + \beta_{D3} \cdot D_{3i} + \beta_E \cdot E_i$$

Donde:

$i = 1, ..., 435$

I_i: Ingreso del i-ésimo cliente.

$$H_{1i} = \begin{cases} 0 & \text{Si el estado residencial del cliente i-ésimo no es "BB"} \\ 1 & \text{Si el estado residencial del cliente i-ésimo es "BB"} \end{cases}$$

$$H_{2i} = \begin{cases} 0 & \text{Si el estado residencial del cliente i-ésimo no es "BC"} \\ 1 & \text{Si el estado residencial del cliente i-ésimo es "BC"} \end{cases}$$

$$H_{3i} = \begin{cases} 0 & \text{Si el estado residencial del cliente i-ésimo no es otro} \\ 1 & \text{Si el estado residencial del cliente i-ésimo es otro} \end{cases}$$

$$Ph_i = \begin{cases} 0 & \text{Si el teléfono fijo proporcionado por el cliente i-ésimo no es válido} \\ 1 & \text{Si el teléfono fijo proporcionado por el cliente i-ésimo es válido} \end{cases}$$

$$Pm_i = \begin{cases} 0 & \text{Si el teléfono móvil proporcionado por el cliente i-ésimo no es válido} \\ 1 & \text{Si el teléfono móvil proporcionado por el cliente i-ésimo es válido} \end{cases}$$

$$C_i = \begin{cases} 0 & \text{Si el cliente i-ésimo no posee otra tarjeta en el mismo banco} \\ 1 & \text{Si el cliente i-ésimo posee otra tarjeta en el mismo banco} \end{cases}$$

L_i: Es el límite del crédito propuesto por el i-ésimo cliente.

$$D_{1i} = \begin{cases} 0 & \text{Si el sistema operativo del dispositivo utilizado por el i-ésimo cliente no es MacOs} \\ 1 & \text{Si el sistema operativo del dispositivo utilizado por el i-ésimo cliente es MacOs} \end{cases}$$

$$D_{2i} = \begin{cases} 0 & \text{Si el sistema operativo del dispositivo utilizado por el i-ésimo cliente no es otro} \\ 1 & \text{Si el sistema operativo del dispositivo utilizado por el i-ésimo cliente es otro} \end{cases}$$

$$D_{3i} = \begin{cases} 0 & \text{Si el sistema operativo del dispositivo utilizado por el i-ésimo cliente no es Windows} \\ 1 & \text{Si el sistema operativo del dispositivo utilizado por el i-ésimo cliente es Windows} \end{cases}$$

E_i: Cantidad de emails registrados en la página del banco del i-ésimo cliente.

Luego, al tener 2 variables continuas (ingreso y credito límite propuesto), se decidió comprobar la linealidad de estas. Para testear esto se implementó el test de razón de verosimilitud, en el que se compara el modelo que considera a la variable como ordinal (asignando scores) y el modelo que considera a la variable como categórica (asignando variables de diseño).

Ingreso

Para crear los modelos se tuvo que categorizar la variable ingreso en 4 categorías con aproximadamente la misma cantidad de observaciones. Las variables creadas son tales que:

Ingreso	I_1	I_2	I_3	S
[0; 0.3]	0	0	0	1
(0.3; 0.6]	1	0	0	2
(0.6; 0.8]	0	1	0	3
(0.8; 0.9]	0	0	1	4

Modelo ordinal:

$$logit(\pi_i) = \beta_0 + \beta_I \cdot S_i + \beta_{H1} \cdot H_{1i} + \beta_{H2} \cdot H_{2i} + \beta_{H3} \cdot H_{3i} + \beta_{Ph} \cdot Ph_i +$$

$$+\beta_{Pm} \cdot Pm_i + \beta_C \cdot C_i + \beta_L \cdot L_i + \beta_{D1} \cdot D_{1i} + \beta_{D2} \cdot D_{2i} + \beta_{D3} \cdot D_{3i} + \beta_E \cdot E_i$$

Modelo categórico:

$$logit(\pi_i) = \beta_0 + \beta_{I_1} \cdot I_{1i} + \beta_{I_2} \cdot I_{2i} + \beta_{I_3} \cdot I_{3i} + \beta_{H1} \cdot H_{1i} + \beta_{H2} \cdot H_{2i} + \beta_{H3} \cdot H_{3i} + \beta_{Ph} \cdot Ph_i +$$

$$+\beta_{Pm} \cdot Pm_i + \beta_C \cdot C_i + \beta_L \cdot L_i + \beta_{D1} \cdot D_{1i} + \beta_{D2} \cdot D_{2i} + \beta_{D3} \cdot D_{3i} + \beta_E \cdot E_i$$

Hipótesis: H_0) La variable es lineal vs H_1) La variable no es lineal

Estadística: $G^2 = -2(Ln(L_{ord}) - Ln(L_{cat})) \underset{H_0}{\sim} \chi^2_2$

Dado que el valor p resulta igual a 0.6906 no se rechaza la hipótesis nula y por lo tanto se introduce la variable al modelo como lineal.

Crédito límite propuesto

Nuevamente, para realizar el test se tuvo que crear variables de diseño y de scores, las cuales son:

Límite cred. propuesto	L_1	L_2	L_3	L_4	S
[0; 200]	0	0	0	0	1
(200; 500]	1	0	0	0	2
(500; 1000]	0	1	0	0	3
(1000; 1500]	0	0	1	0	4
(1500; 2000]	0	0	0	1	5

Modelo ordinal:

$$logit(\pi_i) = \beta_0 + \beta_I \cdot I_i + \beta_{H1} \cdot H_{1i} + \beta_{H2} \cdot H_{2i} + \beta_{H3} \cdot H_{3i} + \beta_{Ph} \cdot Ph_i +$$

$$+\beta_{Pm} \cdot Pm_i + \beta_C \cdot C_i + \beta_L \cdot S_i + \beta_{D1} \cdot D_{1i} + \beta_{D2} \cdot D_{2i} + \beta_{D3} \cdot D_{3i} + \beta_E \cdot E_i$$

Modelo categórico:

$$logit(\pi_i) = \beta_0 + \beta_{I_1} \cdot I_i + \beta_{H1} \cdot H_{1i} + \beta_{H2} \cdot H_{2i} + \beta_{H3} \cdot H_{3i} + \beta_{Ph} \cdot Ph_i +$$

$$+\beta_{Pm} \cdot Pm_i + \beta_C \cdot C_i + \beta_{L_1} \cdot L_{1i} + \beta_{L_2} \cdot L_{2i} + \beta_{L_3} \cdot L_{3i} + \beta_{L_4} \cdot L_{4i} + \beta_{D1} \cdot D_{1i} + \beta_{D2} \cdot D_{2i} + \beta_{D3} \cdot D_{3i} + \beta_E \cdot E_i$$

Hipótesis: H_0) La variable es lineal vs H_1) La variable no es lineal

Estadística: $G^2 = -2(Ln(L_{ord}) - Ln(L_{cat})) \underset{H_0}{\sim} \chi^2_2$

Dado que el valor p resulta igual a 0.1578 no se rechaza la hipótesis nula y por lo tanto se introduce la variable al modelo como lineal.

Además, se propusieron ciertas interacciones, más precisamente la interacción entre la validez del número de teléfono móvil brindado y la tenencia de otra tarjeta en el banco, y entre la validez del número de teléfono móvil brindado y el ingreso. Si bien estas resultaron significativas al realizar el test de razón de verosimilitud, al ya contar con muchas variables se decidió no incluirlas para no complejizar el modelo y su interpretabilidad. Sin embargo este modelo será utilizado más adelante como motivo de comparación con el modelo ajustado.

Una vez definida la componente lineal se decidió comprobar el enlace:

Tabla 3: Test de comprobación de la función de enlace y bondad del ajuste

Enlace	Estadistica test RV	Grados de libertad test RV	Valor p test RV	Estadistica test H-L	Grados de libertad test H-L	Valor p test H-L
Logístico	0.8010	1	0.3708	2.7342	8	0.9499
Probit	1.9706	1	0.1604	5.7644	8	0.6736
Cloglog	0.0015	1	0.9682	2.6860	8	0.9525

Los tres enlaces son apropiados, por lo tanto nos quedamos con el enlace logit por su facilidad en la interpretación.

Modelo estimado

$$logit(\pi_i) = \hat{\beta}_0 + \hat{\beta}_I \cdot I_i + \hat{\beta}_{H1} \cdot H_{1i} + \hat{\beta}_{H2} \cdot H_{2i} + \hat{\beta}_{H3} \cdot H_{3i} + \hat{\beta}_{Ph} \cdot Ph_i +$$
$$+\hat{\beta}_{Pm} \cdot Pm_i + \hat{\beta}_C \cdot C_i + \hat{\beta}_L \cdot L_i + \hat{\beta}_{D1} \cdot D_{1i} + \hat{\beta}_{D2} \cdot D_{2i} + \hat{\beta}_{D3} \cdot D_{3i} + \hat{\beta}_E \cdot E_i$$

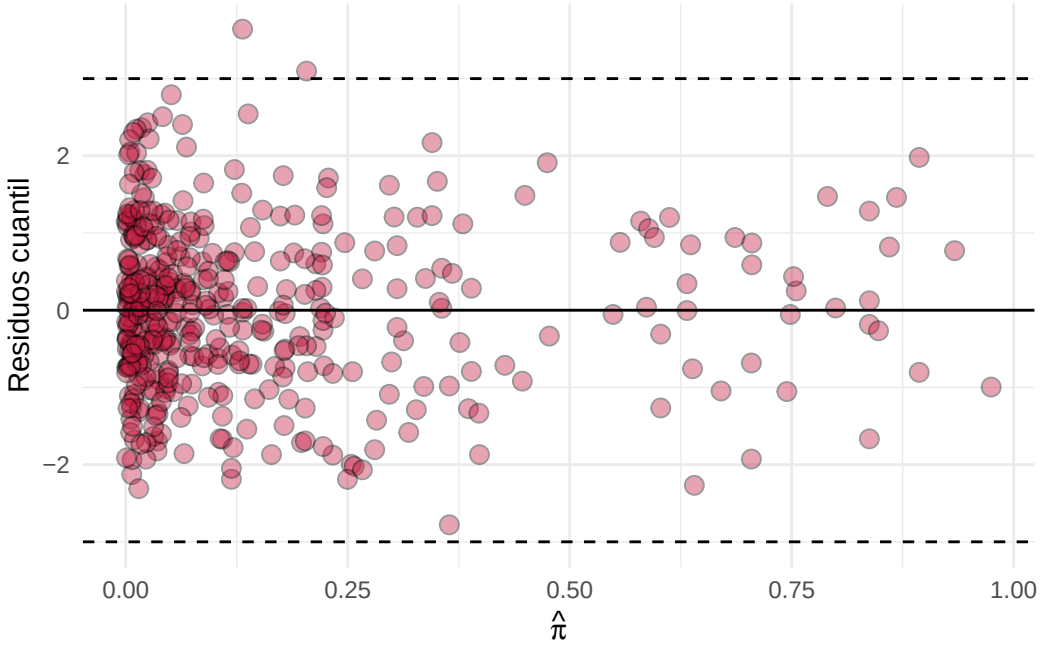
Tabla 4: Coesficientes estimados del modelo ajustado

Beta	Efecto	Coeficiente	LI	LS
$\hat{\beta}_0$	Intercepto	-6.481	-9.119	-3.843
$\hat{\beta}_I$	Ingreso	2.568	1.101	4.035
$\hat{\beta}_{H1}$	Estado residencial:BB	-1.032	-1.938	-0.126
$\hat{\beta}_{H2}$	Estado residencial:BC	-1.465	-2.304	-0.626
$\hat{\beta}_{H3}$	Estado residencial:Otros	-2.582	-4.016	-1.149
$\hat{\beta}_{Ph}$	Teléfono fijo:Válido	-1.328	-2.123	-0.533
$\hat{\beta}_{Pm}$	Teléfono móvil:Válido	-1.015	-1.986	-0.044
$\hat{\beta}_C$	Otra tarjeta:Si	-1.072	-1.991	-0.152
$\hat{\beta}_L$	Crédito límite propuesto	0.001	0.001	0.002
$\hat{\beta}_{D1}$	Sistema operativo:MacOS	1.009	-0.255	2.272
$\hat{\beta}_{D2}$	Sistema operativo:Otro	-0.964	-1.930	0.003
$\hat{\beta}_{D3}$	Sistema operativo:Windows	0.768	-0.046	1.582
$\hat{\beta}_E$	N° de emails distintos (8 sem)	4.408	2.448	6.369

Analisis de residuos

Evaluación de la componente sistemática

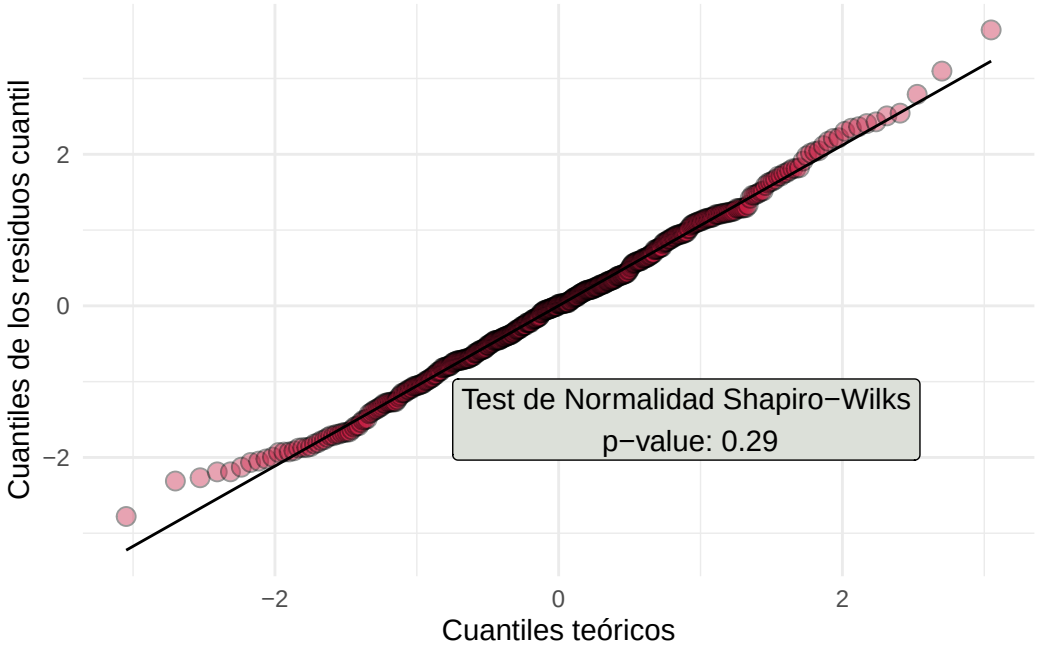
Figura 11: Gráfico de residuos cuantil vs. las probabilidades estimadas



Dado que no se ve ningún patrón y solo 2 puntos se escapan de las bandas se puede decir que la componente sistematica seleccionada es adecuada.

Comprobación de la distribución propuesta

Figura 12: Gráfico probabilístico normal con residuos cuantil



Viendo el gráfico y el test de Shapiro-Wilks para la normalidad de los errores, se puede concluir que la elección de la distribución de la variable es correcta.

Interpretaciones

Tabla 5: Razones de odds del modelo ajustado

RO	Estimación	LI	LS
Límite propuesto	1.1161	1.0518	1.1843
Teléfono fijo:Válido	0.2650	0.1197	0.5868
Ingreso	1.2928	1.1164	1.4971
Tenencia otra tarjeta	0.3425	0.1366	0.8589

La chance de que un cliente cometa fraude aumenta entre un 5% y un 18% cuando el límite del crédito propuesto aumenta en 100 unidades monetarias, cuando las demás variables son constantes.

A su vez, cuando el cliente proporciona un telefono fijo válido, la chance de que este cometa fraude es como mínimo un 41% y como máximo un 88% menor que la de una persona que no proporcionó un telefono fijo válido, cuando el resto de las variables son constantes.

Cuando el ingreso del cliente aumenta en un decil, la chance de que cometa fraude aumenta entre un 12% y un 50%, cuando las variables son constantes.

Los clientes que tienen otra tarjeta en el mismo banco tienen una chance de cometer fraude entre un 14% y un 86% menor que aquellos que no tienen, cuando el resto de las variables son constantes.

Las características de las personas más propensas a cometer fraude son:

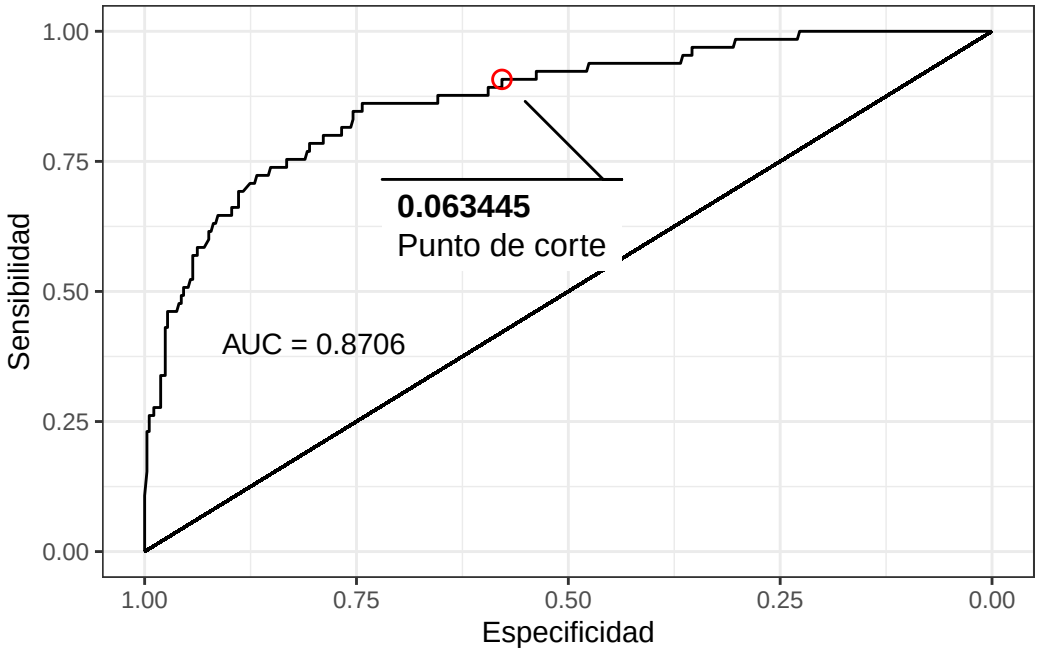
- Ingreso alto
- Estado residencial del aplicante “BA”
- Número de teléfono fijo proporcionado no válido
- Número de celular proporcionado no válido
- No posee otra tarjeta en el mismo banco
- Límite del crédito propuesto alto
- Sistema operativo usado MacOS
- 2 e-mails registrados en la página del banco en las últimas 8 semanas

Las características de las personas menos propensas a cometer fraude son:

- Ingreso bajo
- Estado residencial del aplicante otro (distinto de “BA”, “BB”, “BC”)
- Número de teléfono fijo proporcionado válido
- Número de celular proporcionado válido
- Posee otra tarjeta en el mismo banco
- Límite del crédito propuesto bajo
- Sistema operativo usado otros (distintos de Windows, MacOS y Linux)
- Un solo e-mail registrado en la página del banco en las últimas 8 semanas

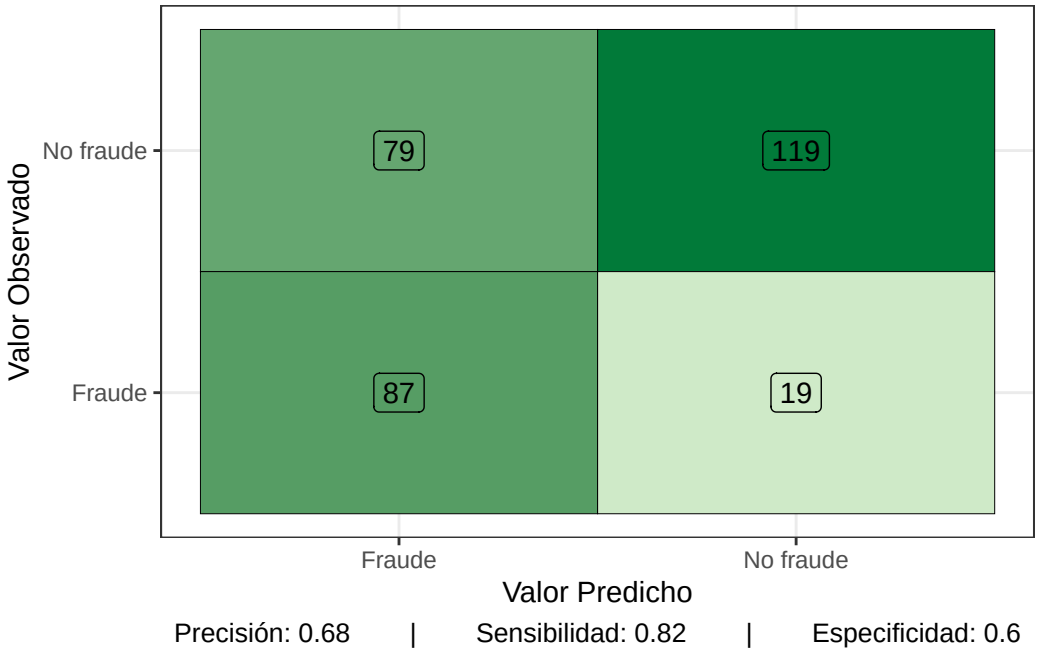
Capacidad predictiva

Figura 13: Curva ROC del modelo ajustado



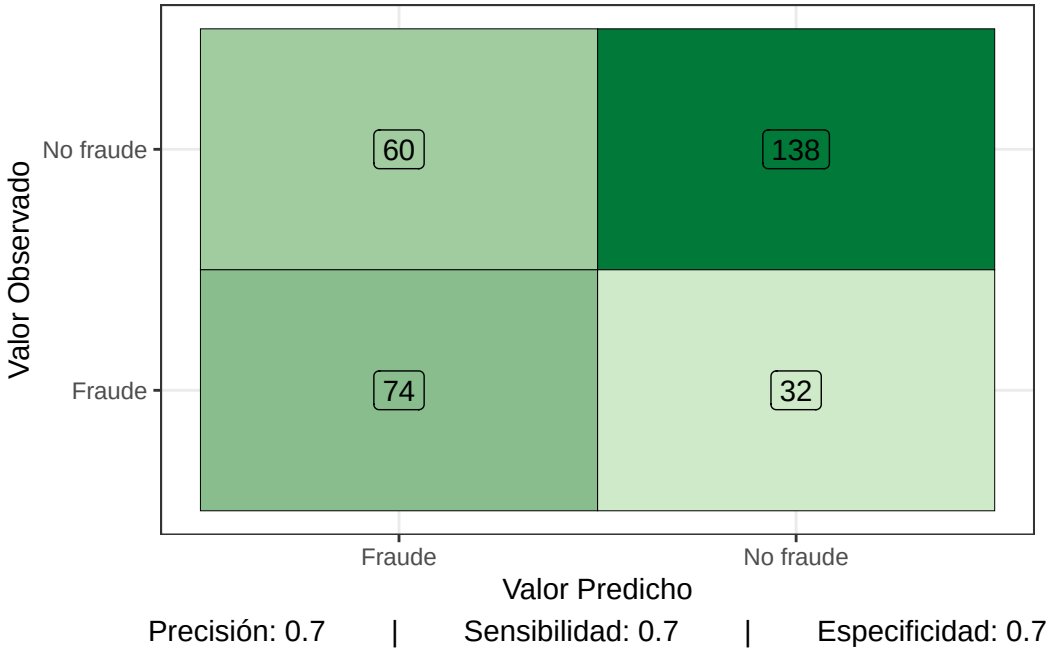
Se decidió elegir el mejor punto de corte como aquel que maximiza la especificidad (probabilidad de que un no fraude sea clasificado como tal) garantizando una sensibilidad (probabilidad de que un fraude sea clasificado como tal) de al menos un 90%, esto para evitar el máximo número de fraudes posibles sin afectar a los clientes legítimos.

Figura 14: Matriz de confusión del modelo ajustado ante nuevos clientes



Además, se quiso evaluar la capacidad predictiva del mismo modelo agregando las 2 interacciones significativas mencionadas en el ajuste del modelo:

Figura 15: Matriz de confusión del modelo con interacciones ante nuevos clientes



Observando las matrices de confusión y las métricas de comparación, haciendo énfasis en la sensibilidad,

se puede notar que el modelo de efectos principales tiene una mejor capacidad preditiva en base al objetivo propuesto que el modelo que tiene 2 efectos más.

Discusión

Los resultados obtenidos en este informe permitirán a los bancos tomar medidas para detectar las transacciones fraudulentas con una mayor precisión, impactando de manera positiva en la seguridad y balance financiero de los bancos, generando así una mayor confianza para sus clientes.

Si bien los modelos lineales generalizados empleados ayudaron a resolver la problemática, se debe tener en cuenta que para este tipo de escenarios, otras técnicas de apredizaje supervisado más enfocadas a la predicción (k-vecinos más cercanos, redes nueronales, entre otros), podrían ser más adecuadas, a cambio de sacrificar interpretabilidad.

La metodología y código utilizado pueden ser consultados en el repositorio dedicado al trabajo.