LICENCIATURA EN ESTADÍSTICA

Pradicción 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

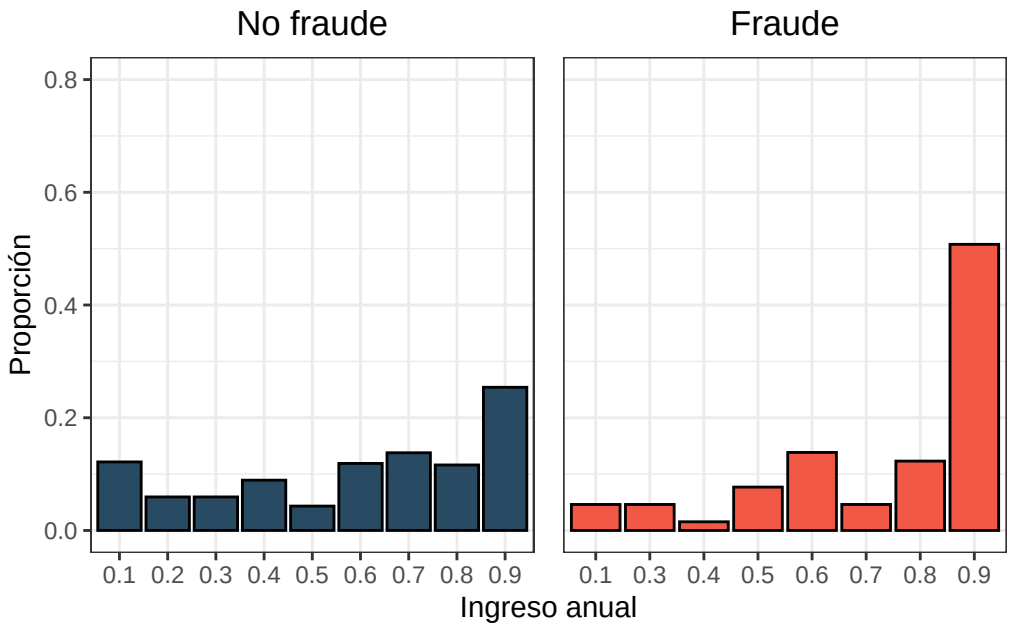
El fraude con tarjetas de crédito es una de las principales amenazas que sufren los bancos. Con el auge de la tecnología las transacciones digitales facilitaron los traspasos de dinero y los medios de pago electrónicos son algo de cada día, pero junto con las ventajas también vinieron las consecuencias, 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 mismo pais que el banco
- `device_os`: Sistema operativo del dispositivo que 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

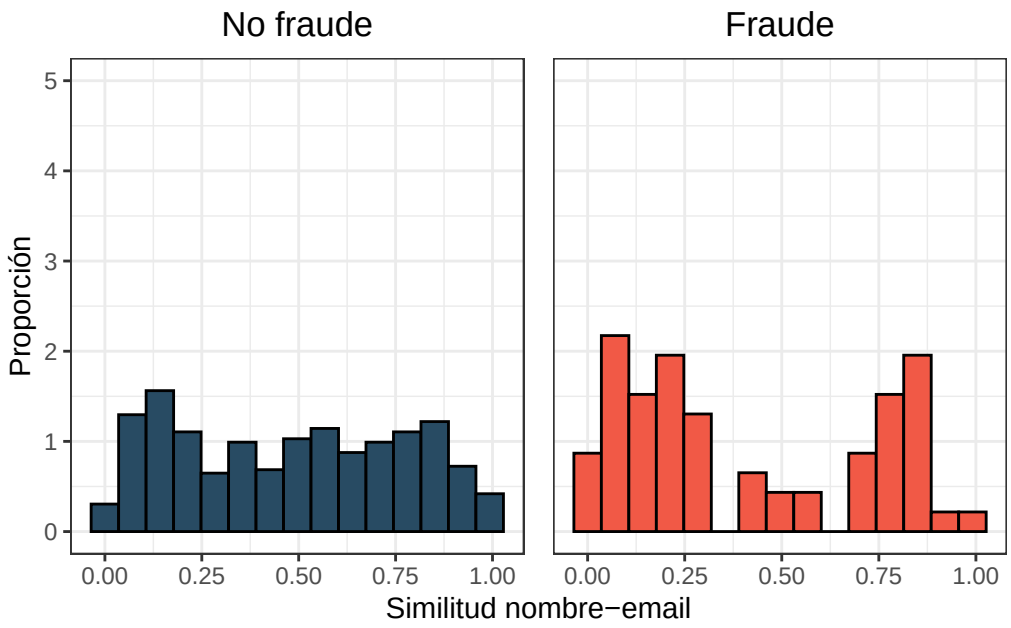
Análisis descriptivo

Figura 1: Distribución del ingreso anual según si la transacción es fraudulenta o no



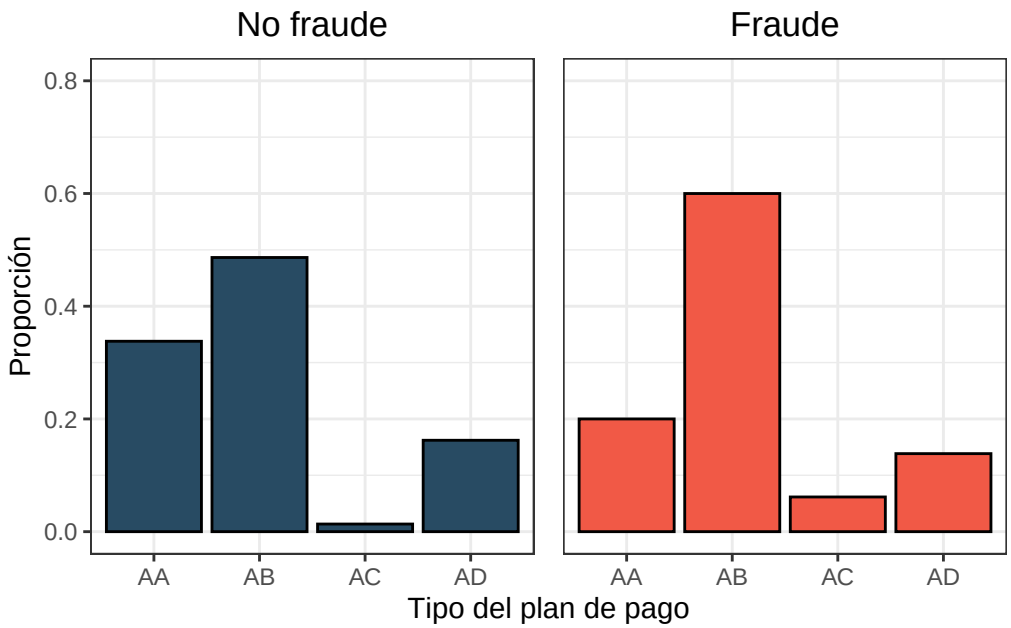
Se puede observar que las personas que cometieron fraude tienden a tener un ingreso anual registrado mayor. La distribución tiene una mayor asimetría a la izquierda.

Figura 2: Distribución del índice de similitud entre en nombre del solicitante y el nombre en el email según si la transacción es fraudulenta o no



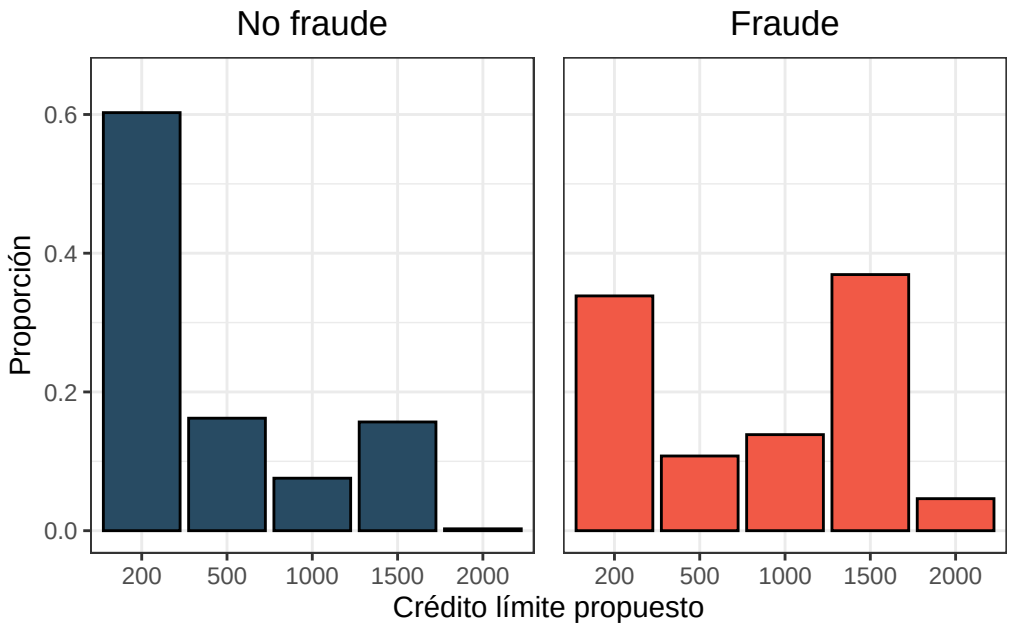
Las transacciones realizadas con emails no muy similares al nombre real de la persona parecen ser más propensas a ser fraudulentas.

Figura 3: Proporción del tipo de pago según si la transacción es fraudulenta o no



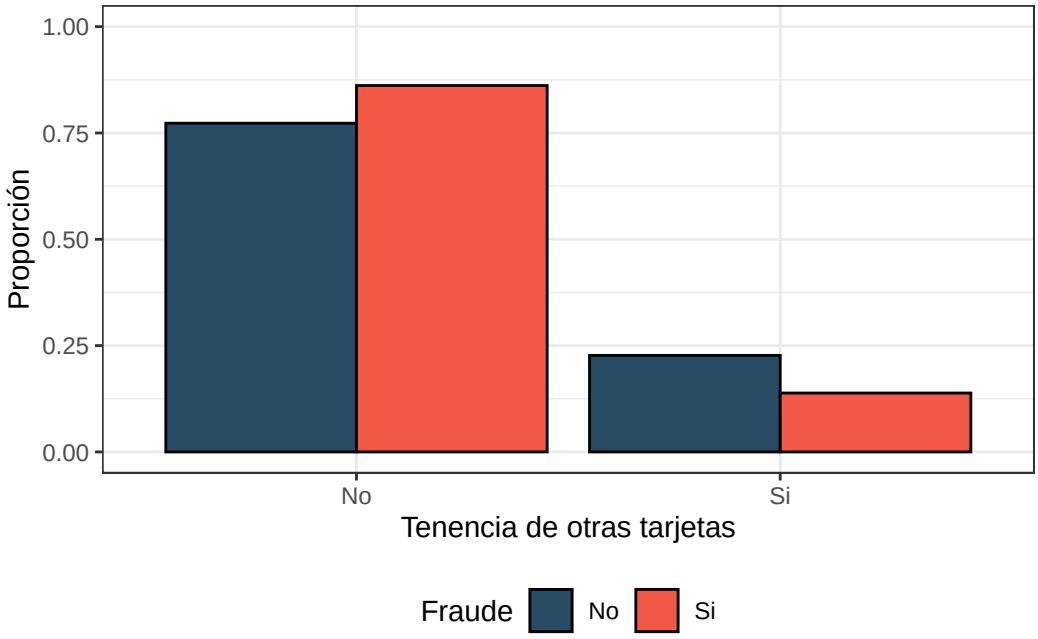
En general, las personas que cometen fraude parecen preferir los métodos de pago “AB” y “AC” por encima del resto, al contrario de las personas que operan de manera legítima que prefieren de igual manera los tipos de pago “AA”, “AB” y “AC”. Se puede notar también que la forma de pago “AE” no es muy popular.

Figura 4: Distribución del límite crediticio propuesto por el solicitante según si la transacción es fraudulenta o no



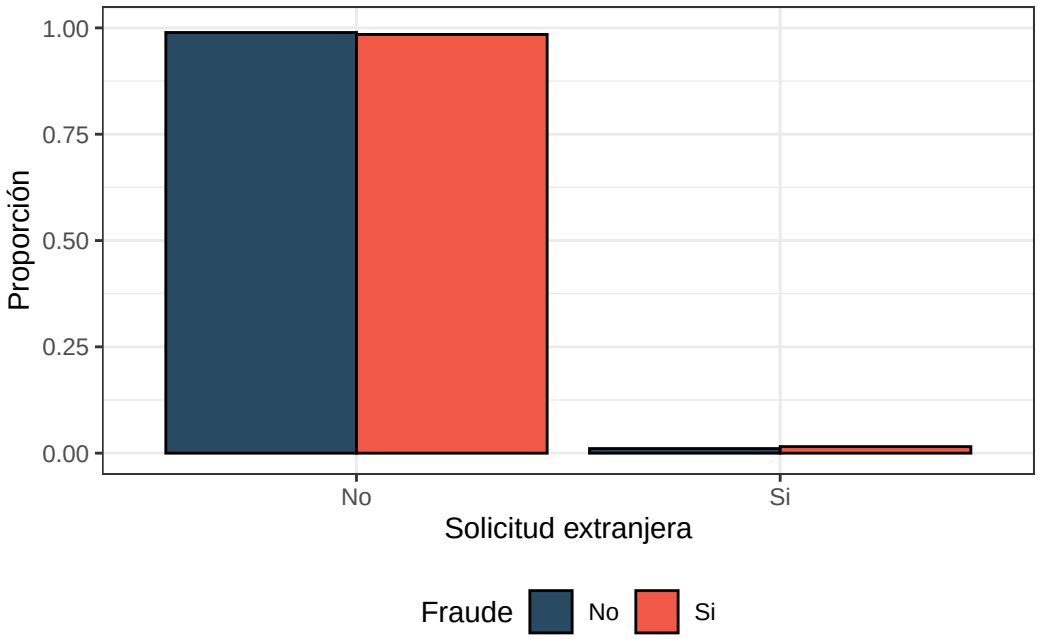
Se puede destacar en este gráfico que las personas que cometen fraude son ligeramente más propensas a pedir créditos más altos.

Figura 5: Proporción de la tenencia de otra tarjeta en el mismo banco según si la transacción es fraudulenta o no



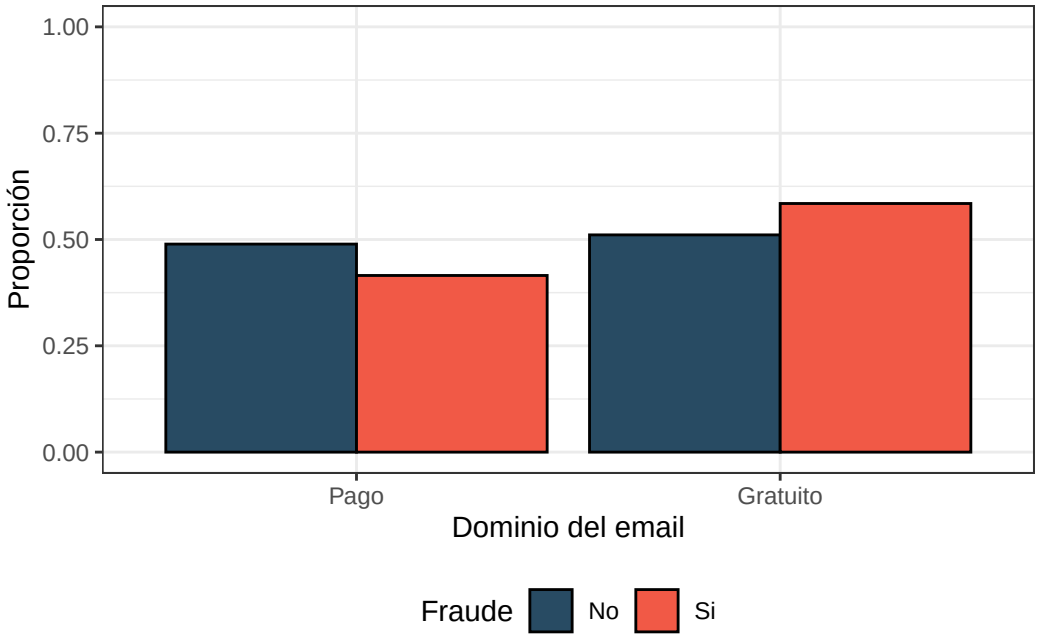
En cuanto a la tenencia de otra tarjeta en el mismo banco, suele ser común no poseer otra, sin embargo las personas que operan de forma legal se inclinan a tener más de una tarjeta un poco más que aquellos que cometen fraude.

Figura 6: Proporción de la locación de la solicitud según si la transacción es fraudulenta o no



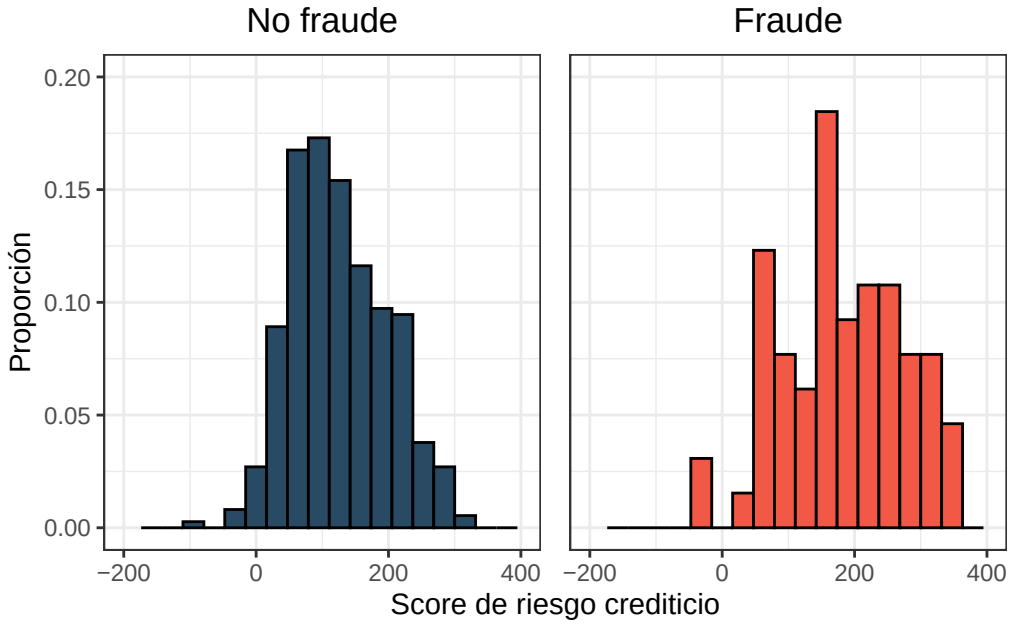
Se puede observar que las personas que cometen fraude, parecen hacer más solicitudes del exterior que las personas que no cometen fraude, aunque la diferencia parece ser sutil.

Figura 7: Proporción del tipo de dominio del email según si la transacción es fraudulenta o no



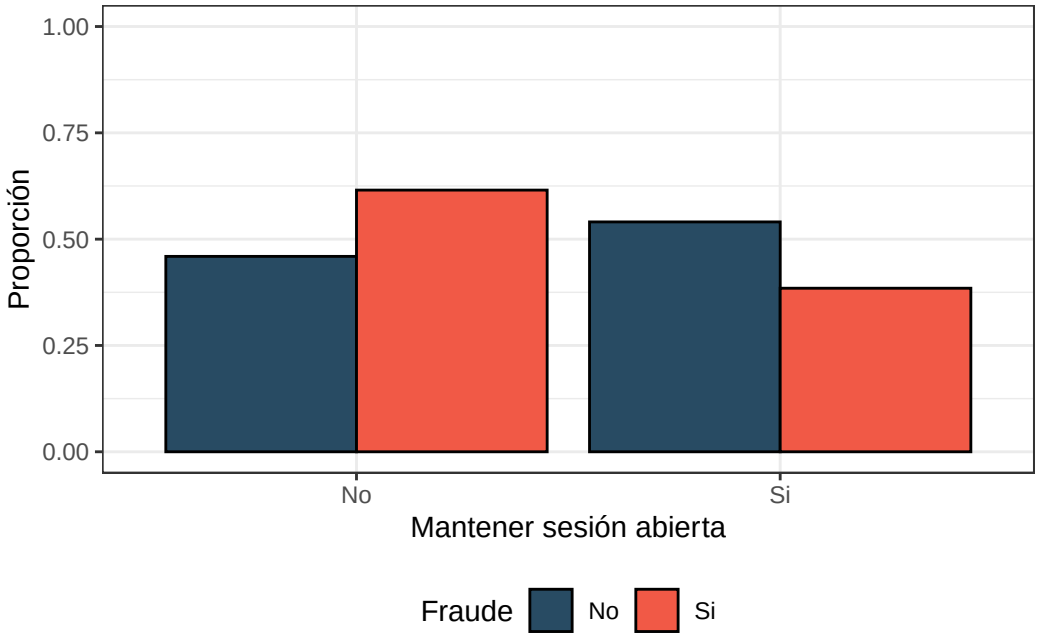
También se puede destacar que las operaciones fraudulentas parecen ser más comunes cuando el dominio del email del solicitante es gratuito que cuando es pago.

Figura 8: Distribución del score de riesgo interno según si la transacción es fraudulenta o no



La distribución del score de riesgo para las personas que cometen fraude es simétrica y centrada alrededor de 200, mientras que la distribución del score de riesgo para las personas que no cometen fraude parece ser más asimétrica y tener una media menor.

Figura 9: 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:

$$logit(\pi_i) = \beta_0 + \beta_I \cdot I_i + \beta_{H1} \cdot H_{1i} + \beta_{H2} \cdot H_{2i} + \beta_{H3} \cdot H_{3i} + \beta_{H4} \cdot H_{4i} + \beta_{H5} \cdot H_{5i} + \beta_{Ph} \cdot Ph_i +$$

$$+\beta_{Pm} \cdot Pm_i + \beta_C \cdot C_i + \beta_L \cdot L_i + \beta_{D1} \cdot D_{1i} + \beta_{D2} \cdot D_{2i} + \beta_{D3} \cdot D_{3i} + \beta_{D4} \cdot D_{4i} + \beta_E \cdot E_i$$

Luego, al tener 2 variables continuas, se decidió comprobar la linealidad de estas:

Además, se propucieron ciertas interacciones:

Sin embargo se mantuvo el modelo lineal presentado anteriormente (AGREGAR REFERENCIA A LA ECUACION)

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

Tabla 1: Test de comprabación de la función de enlace y bondad de 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	0.2831	1	0.3799	3.4613	8	0.9022
Probit	0.4987	1	0.1689	5.9454	8	0.6534
Cloglog	0.0115	1	0.9663	3.1135	8	0.9270

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

Modelo estimado

$$logit(\pi_i) = \hat{\beta}_0 + \hat{\beta}_I \cdot I_i + \hat{\beta}_{H1} \cdot H_{1i} + \hat{\beta}_{H2} \cdot H_{2i} + \hat{\beta}_{H3} \cdot H_{3i} + \hat{\beta}_{H4} \cdot H_{4i} + \hat{\beta}_{H5} \cdot H_{5i} + \hat{\beta}_{Ph} \cdot Ph_i +$$
$$+\hat{\beta}_{Pm} \cdot Pm_i + \hat{\beta}_C \cdot C_i + \hat{\beta}_L \cdot L_i + \hat{\beta}_{D1} \cdot D_{1i} + \hat{\beta}_{D2} \cdot D_{2i} + \hat{\beta}_{D3} \cdot D_{3i} + \hat{\beta}_{D4} \cdot D_{4i} + \hat{\beta}_E \cdot E_i + \hat{\beta}_{IPm} \cdot I_i \cdot Pm_i + \hat{\beta}_{CPm} \cdot C_i \cdot Pm_i$$

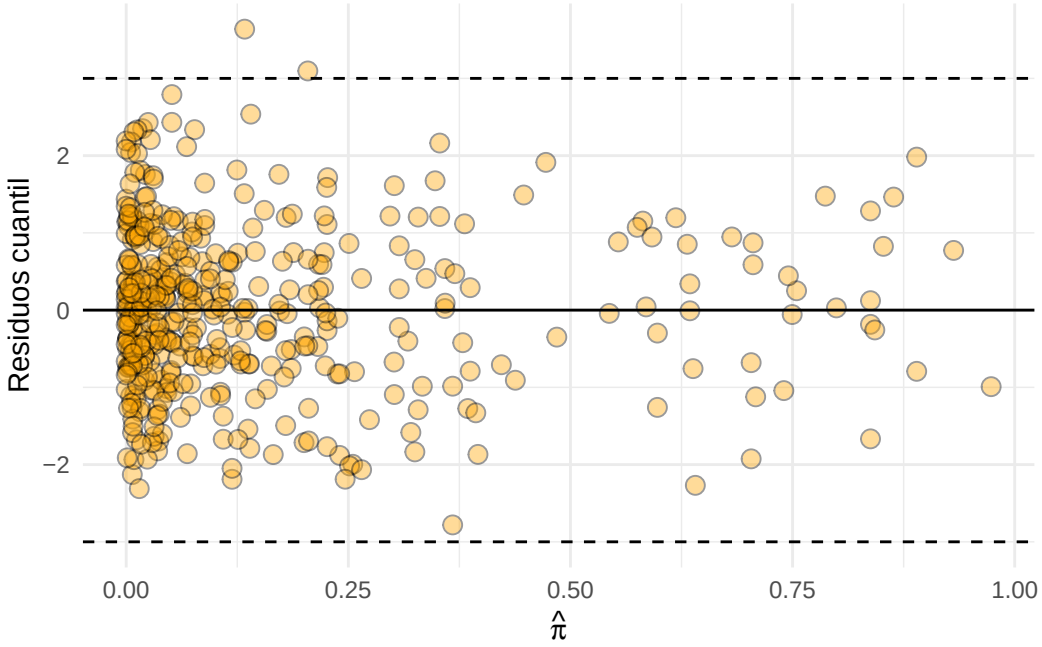
Tabla 2: Coesficientes estimados del modelo ajustado

Beta	Efecto	Coeficiente	LI	LS
$\hat{\beta}_0$	Intercepto	-6.466	-9.099	-3.832
$\hat{\beta}_I$	Ingreso	2.592	1.128	4.056
$\hat{\beta}_{H1}$	Estado residencial:BB	-1.037	-1.942	-0.132
$\hat{\beta}_{H2}$	Estado residencial:BC	-1.455	-2.292	-0.617
$\hat{\beta}_{H3}$	Estado residencial:BD	-16.646	-2091.365	2058.072
$\hat{\beta}_{H4}$	Estado residencial:BE	-2.373	-3.836	-0.909
$\hat{\beta}_{H5}$	Estado residencial:BF	-16.237	-5474.750	5442.276
$\hat{\beta}_{Ph}$	Teléfono fijo:Válido	-1.318	-2.113	-0.524
$\hat{\beta}_{Pm}$	Teléfono móvil:Válido	-0.983	-1.952	-0.014
$\hat{\beta}_C$	Otra tarjeta:Si	-1.075	-1.991	-0.158
$\hat{\beta}_L$	Límite propuesto	0.001	0.000	0.002
$\hat{\beta}_{D1}$	Sistema operativo:MacOS	0.978	-0.283	2.238
$\hat{\beta}_{D2}$	Sistema operativo:Otro	-0.935	-1.904	0.035
$\hat{\beta}_{D3}$	Sistema operativo:Windows	0.768	-0.047	1.582
$\hat{\beta}_{D4}$	Sistema operativo:OSx11	-15.595	-4419.005	4387.816
$\hat{\beta}_E$	N° de emails distintos (8 sem)	4.355	2.399	6.311

Analisis de residuos

Evaluación de la componente sistemática

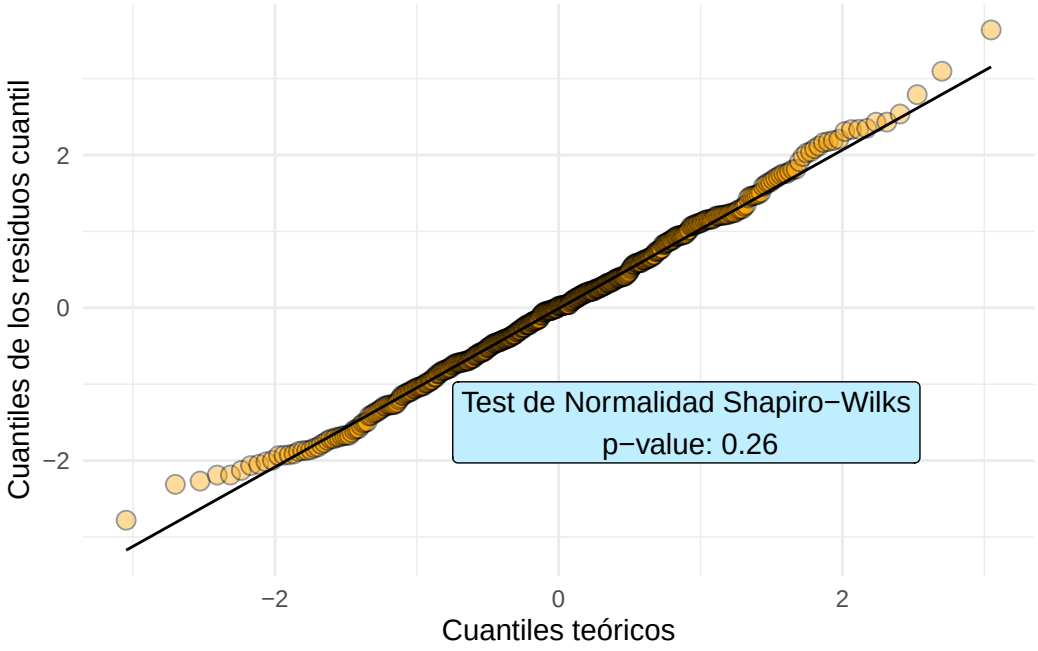
Figura 10: Gráfico de residuos cuantil vs. las probabilidades estimadas



Dado que no se ve ningún patrón y ningun punto se escapa de las bandas se puede decir que la componente sistematica seleccionada es adecuada.

Comprobación de la distribución propuesta

Figura 11: 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 3: Razones de odds del modelo ajustado

RO	Estimacion	LI	LS
Límite propuesto	1.1151	1.0510	1.1832
Teléfono Fijo:Válido	0.2675	0.1209	0.5919
Ingreso	1.2959	1.1194	1.5003
Tenencia otra tarjeta	0.3414	0.1366	0.8536

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 15% 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:

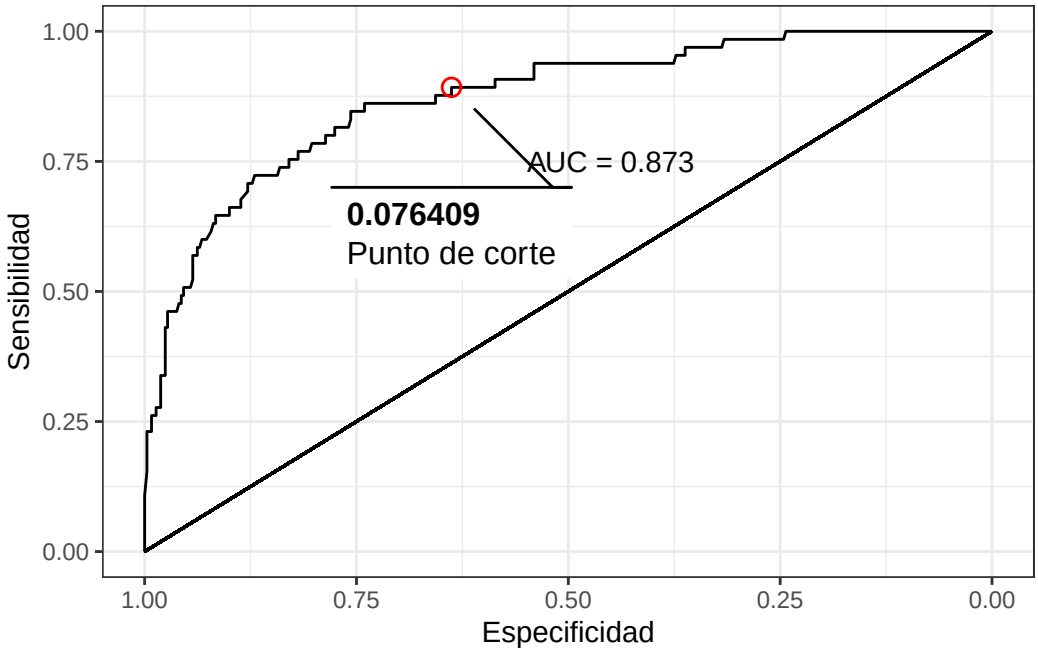
- Ingreso alto
- Estado residencial del aplicante “BA”
- Número de teléfono fijo proporcionado no válido
- Número de celular proporcionado no válido
- No posee otra tarjeta en el mismo banco
- Límite del crédito propuesto alto
- Sistema operativo usado macOS
- 2 emails registados en la página del banco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

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

- Ingreso Bajo
- Estado residencial del aplicante “BE”
- 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
- Un solo email registado en la página del banco en las últimas 8 semanas

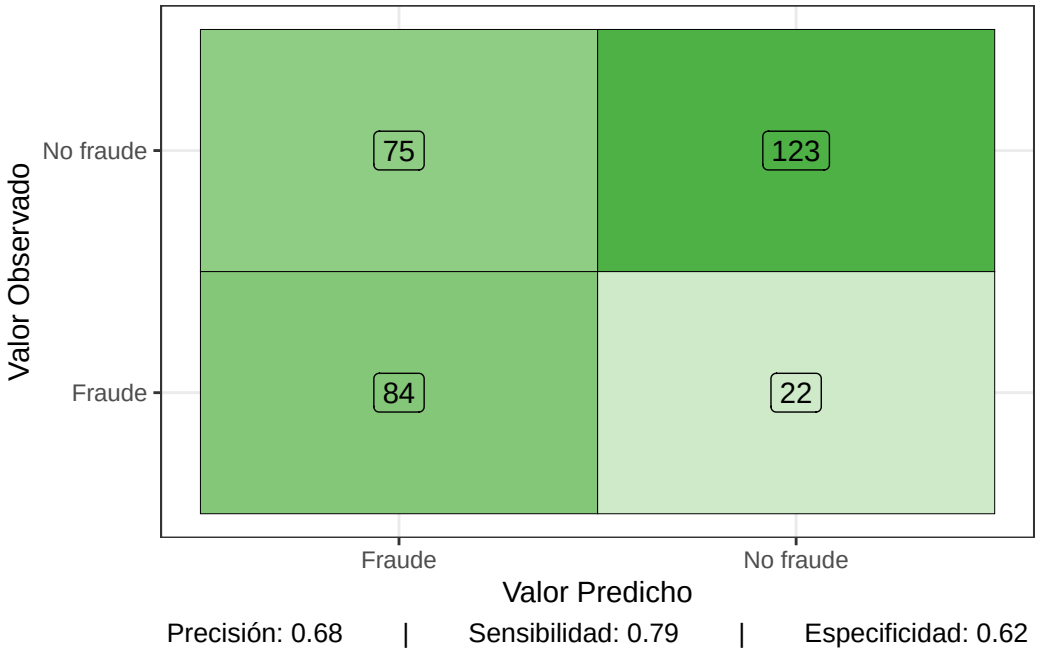
Capacidad predictiva

Figura 12: Curva ROC del modelo ajustado



Se decidió elegir el mejor punto de corte como aquel que maximiza la especificidad garantizando una sensibilidad de al menos un 90%, esto para evitar el máximo número de fraudes posibles sin afectar a los clientes legítimos.

Figura 13: Matriz de confusión del modelo ajustado ante nuevos clientes



Además, se quiso evaluar la capacidad predictiva del mismo modelo agregando 2 interacciones:

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. Esto puede adjudicarse al sobreajuste que el modelo tiene sobre los datos que se usaron como entrenamiento, ya que si bien un modelo con muchos efectos va a predecir mejor sobre los datos que modela, ante nuevos datos el ajuste no será tan bueno.