Universidad Nacional de Rosario

LICENCIATURA EN ESTADÍSTICA

Guardianes de la galaxia financiera

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.

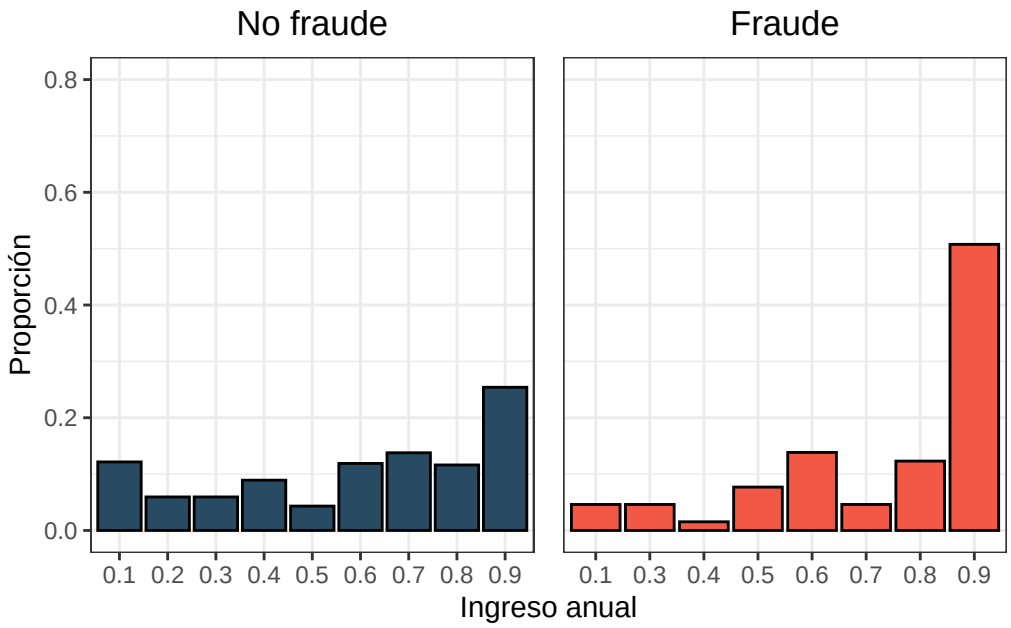
Variables:

- `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
- `prev_address_months_count`: Es el número de meses que la persona estuvo viviendo en su locacion anterior
- `current_address_months_count`: Es el número de meses que la persona estuvo viviendo en su locacion actual
- `customer_age`: Edad del cliente en décadas
- `days_since_request`: Días desde la solicitud
- `intended_balcon_amount`: Valor de la transferecia inicial para aplicar al credito
- `payment_type`: Tipo del plan de pago
- `zip_count_4w`: Número de aplicaciónes con el mismo código postal en las últimas 4 semanas
- `velocity_6h`: Es la velocidad del total de solicitudes de transferencias de la tarjeta en las últimas 6 horas
- `velocity_24h`: Es la velocidad del total de solicitudes de transferencias de la tarjeta en las últimas 24 horas
- `velocity_4w`: Es la velocidad del total de solicitudes de transferencias de la tarjeta en las últimas 4 semanas
- `bank_branch_count_8w`: Número total de solicitudes en la seleccionada rama del banco en las últimas 8 semanas
- `date_of_birth_distinct_emails_4w`: Número de emails de aplicantes con la misma fecha de nacimiento en las últimas 4 semanas
- `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
- `bank_months_count`: Antiguedad de la cuenta anterior en meses
- `has_other_cards`: Indicador 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
- `source`: Fuente online de la aplicación (Internet / app movil)
- `session_length_in_minutes`: Tiempo de la sesion en la pagina del banco en minutos
- `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
- `device_fraud_count`: Número de solicitudes fraudulentas desde el dispositivo utilizado
- `month`: Mes en el que fue realizada la solicitud

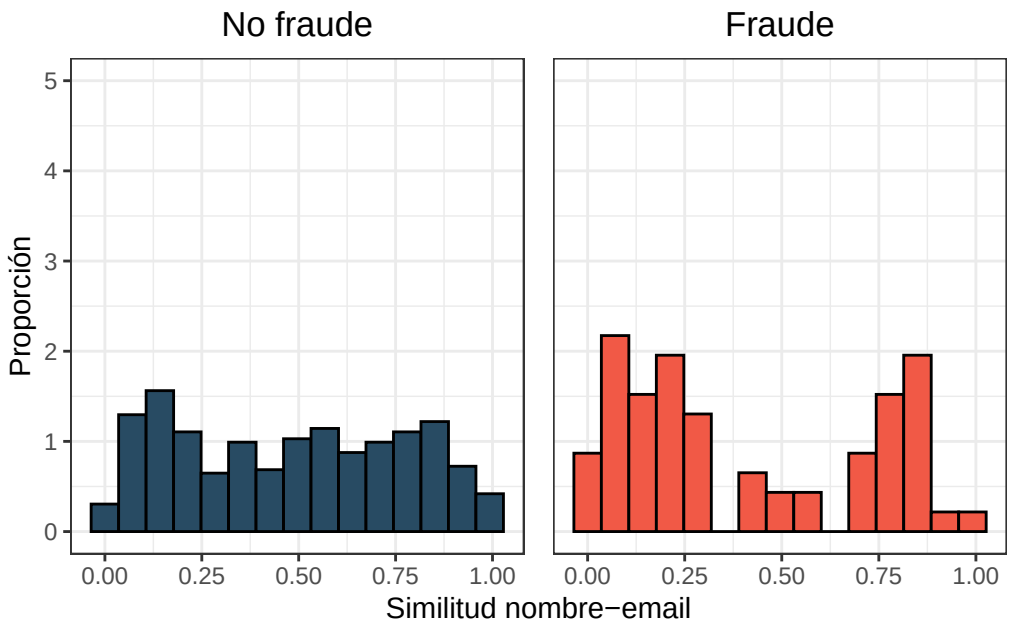
Análisis descriptivo

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



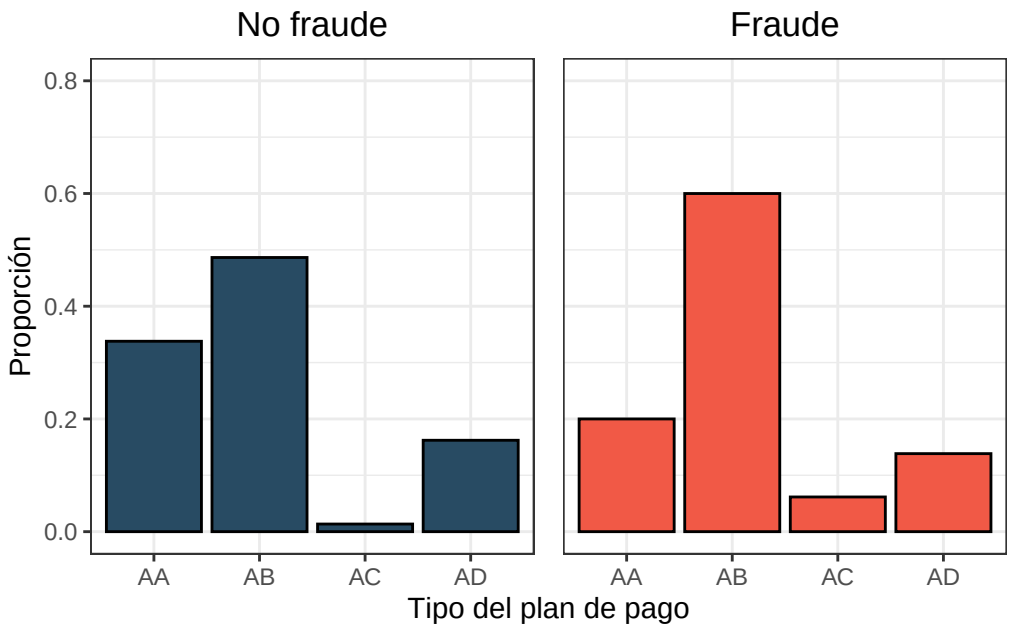
Se puede observar que las personas que cometieron fraude tienden a tener un ingreso anual registrado mayor. La distribución tiene una mayor asimetría a la izquierda.

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



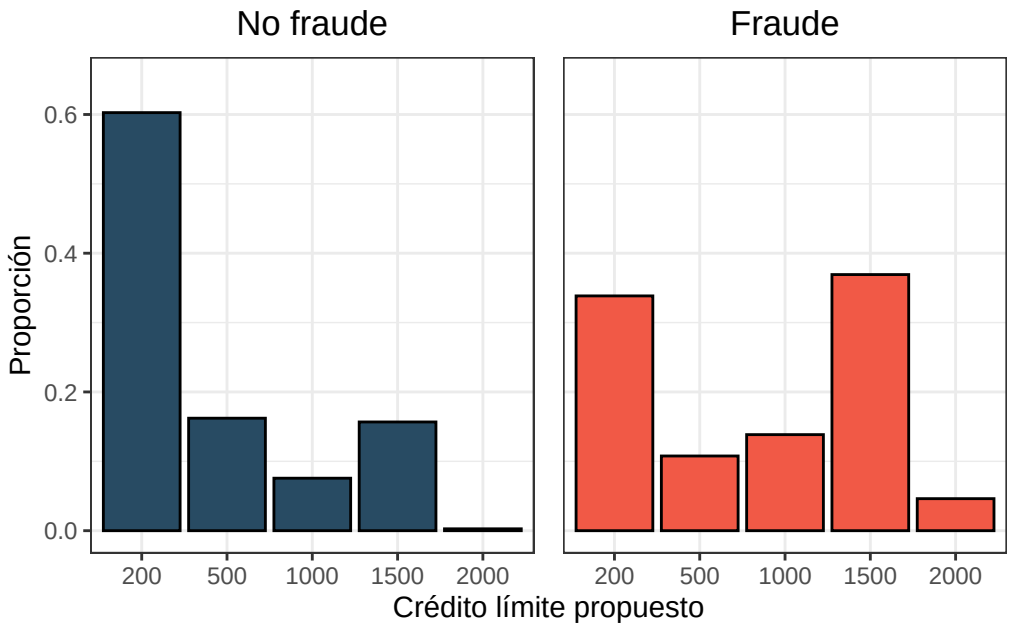
Las transacciones realizadas con emails no muy similares al nombre real de la persona parecen ser más propensas a ser fraudulentas.

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



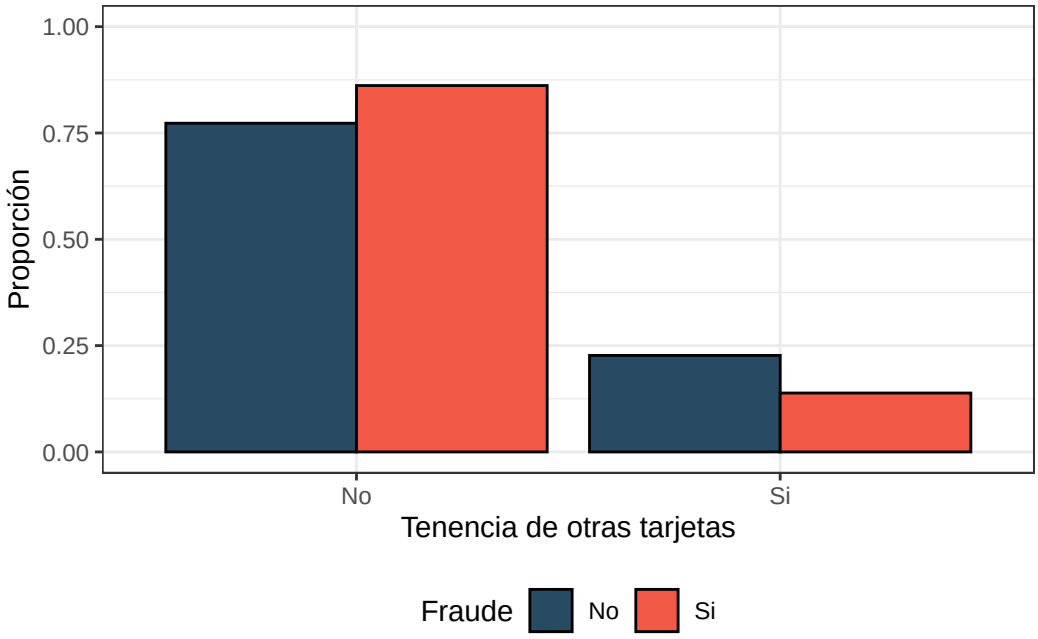
En general, las personas que cometen fraude parecen preferir los métodos de pago “AB” y “AC” por encima del resto, al contrario de las personas que operan de manera legítima que prefieren de igual manera los tipos de pago “AA”, “AB” y “AC”. Se puede notar también que la forma de pago “AE” no es muy popular.

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



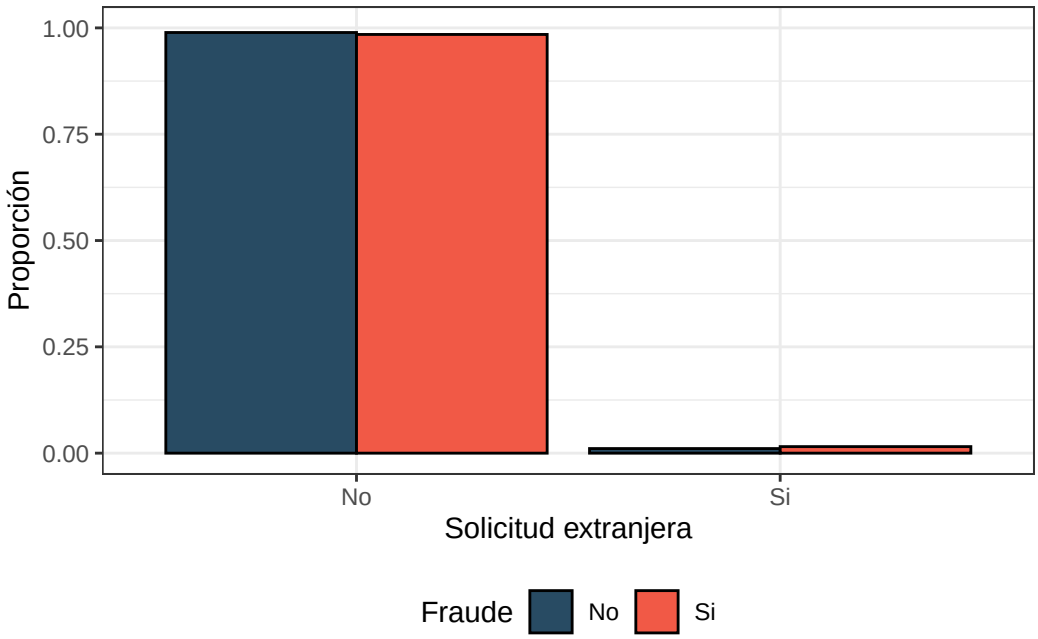
Se puede destacar en este gráfico que las personas que cometen fraude son ligeramente más propensas a pedir créditos más altos.

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



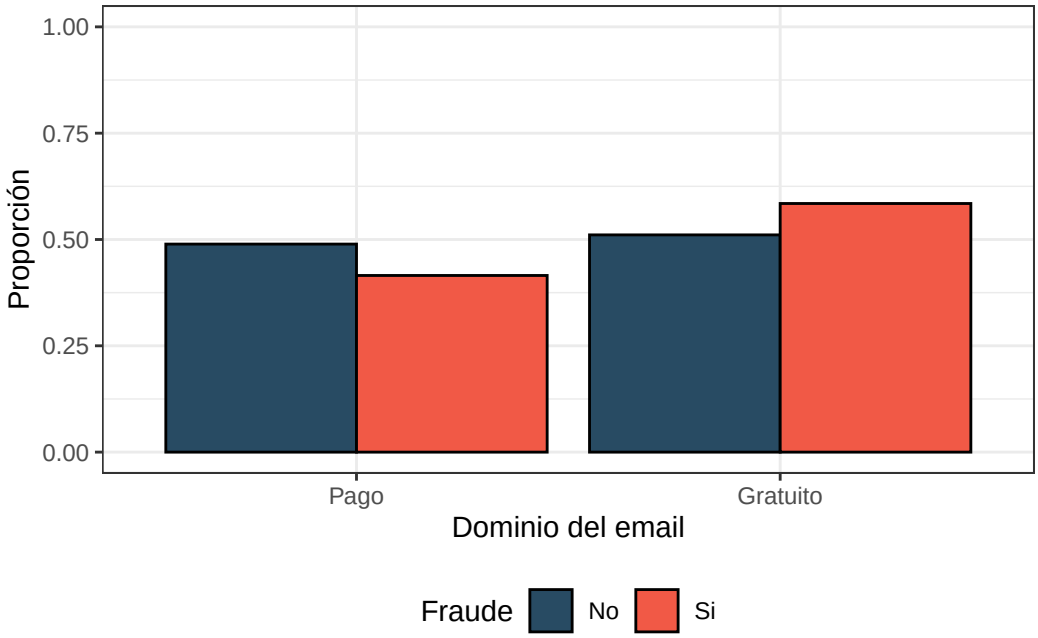
En cuanto a la tenencia de otra tarjeta en el mismo banco, suele ser común no poseer otra, sin embargo las personas que operan de forma legal se inclinan a tener más de una tarjeta un poco más que aquellos que cometen fraude.

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



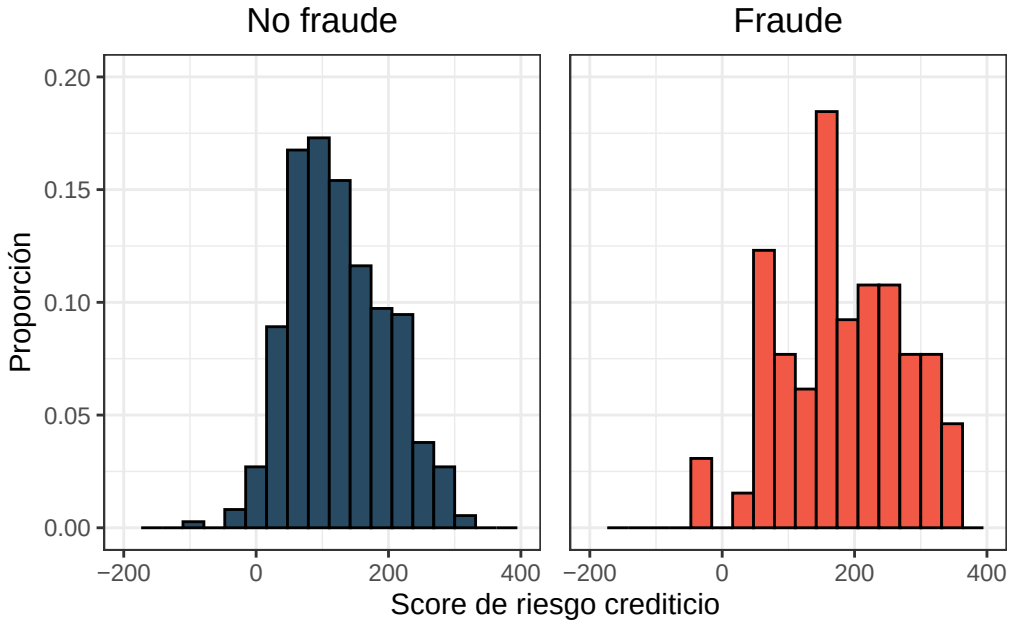
Se puede observar que las personas que cometen fraude, parecen hacer más solicitudes del exterior que las personas que no cometen fraude, aunque la diferencia parece ser sutil.

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



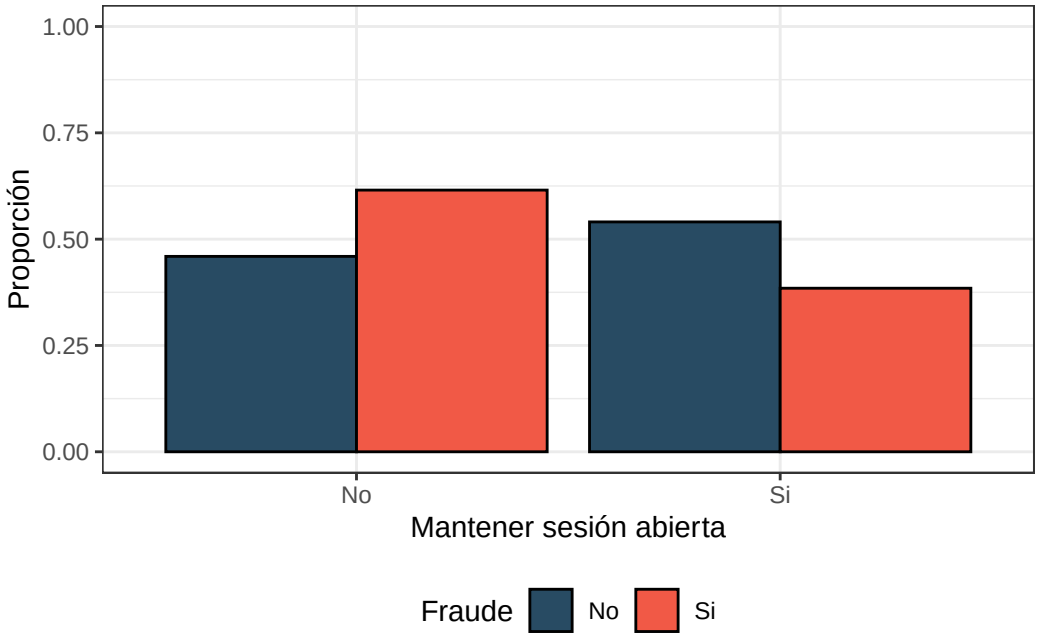
También se puede destacar que las operaciones fraudulentas parecen ser más comunes cuando el dominio del email del solicitante es gratuito que cuando es pago.

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:

$$logit(\pi_i) = \beta_0+\beta_I\cdot I_i+\beta_{H1}\cdot H_{1i}+\beta_{H2}\cdot H_{2i}+\beta_{H3}\cdot H_{3i}+\beta_{H4}\cdot H_{4i}+\beta_{H5}\cdot H_{5i}+\beta_{Ph}\cdot Ph_i+\beta_{Pm}\cdot Pm_i+\beta_C\cdot C_i+\beta_L\cdot L_i+\beta_D\cdot D_i+($$

Se procede a comprobar el enlace:

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

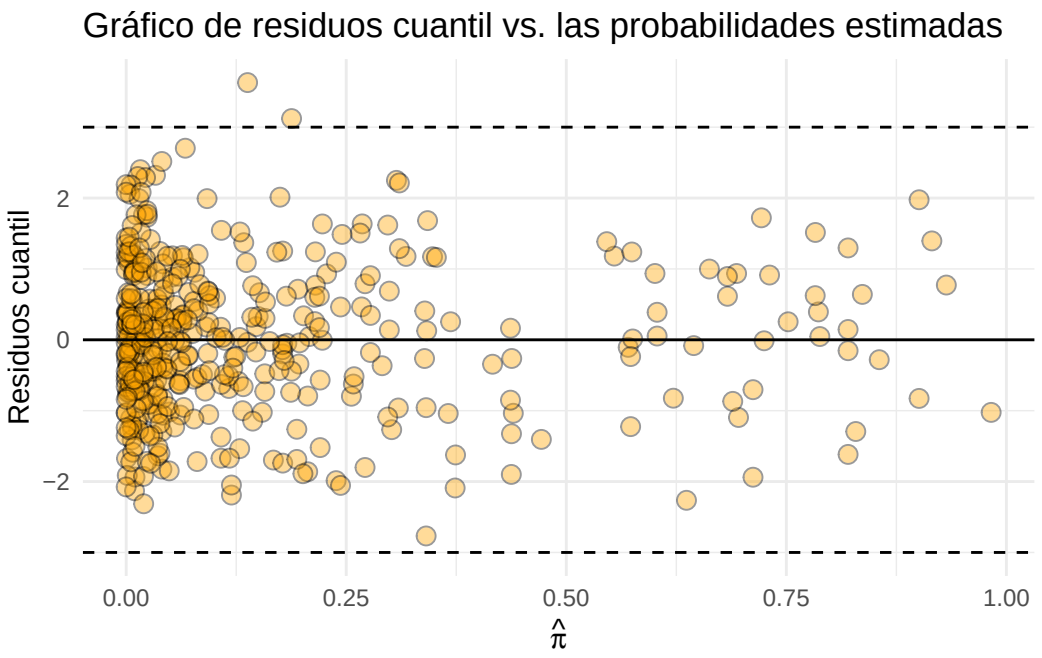
Modelo estimado

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

- $\hat{\beta}_0 = -10.48$
- $\hat{\beta}_I = 8.03$
- $\hat{\beta}_{H1} = -1.05$
- $\hat{\beta}_{H2} = -1.62$
- $\hat{\beta}_{H3} = -17.52$
- $\hat{\beta}_{H4} = -2.5$
- $\hat{\beta}_{H5} = -17.16$
- $\hat{\beta}_{Ph} = -1.31$
- $\hat{\beta}_{Pm} = 3.21$
- $\hat{\beta}_C = -18.56$
- $\hat{\beta}_L = 0.001$
- $\hat{\beta}_{D1} = 1.15$
- $\hat{\beta}_{D2} = -1.02$
- $\hat{\beta}_{D3} = 0.75$
- $\hat{\beta}_{D4} = -16.42$
- $\hat{\beta}_E = 4.55$
- $\hat{\beta}_{IPm} = -5.99$
- $\hat{\beta}_{CPm} = 17.84$

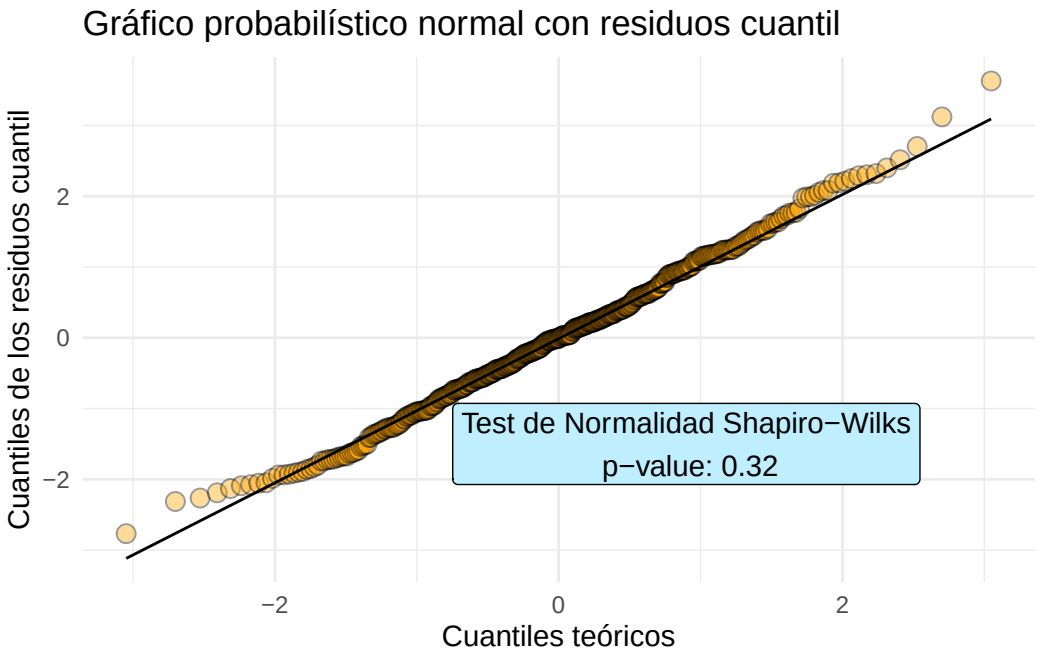
Analisis de residuos

Evaluación de la componente sistemática



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



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

Efecto	Coeficiente	LI	LS
Intercepto	-10.482	-16.490	-4.473
Ingreso	8.025	1.711	14.340
Estado residencial:BB	-1.046	-1.963	-0.130
Estado residencial:BC	-1.620	-2.478	-0.762
Estado residencial:BD	-17.520	-3438.635	3403.594
Estado residencial:BE	-2.505	-4.010	-1.000
Estado residencial:BF	-17.156	-9032.224	8997.912
Teléfono Fijo:Válido	-1.310	-2.112	-0.509
Teléfono Móvil:Válido	3.215	-2.032	8.462
Otra tarjeta:Si	-18.556	-3371.347	3334.235
Límite propuesto	0.001	0.000	0.002
Sistema Operativo:MacOS	1.147	-0.136	2.429
Sistema Operativo:Otro	-1.017	-2.014	-0.021
Sistema Operativo:Windows	0.749	-0.084	1.582
Sistema Operativo:OSx11	-16.423	-7272.770	7239.924
N° de Emails Distintos_8w	4.553	2.568	6.537
Ingreso::Teléfono Móvil:Válido	-5.987	-12.413	0.439
Otra tarjeta:Si::Teléfono Móvil:Válido	17.835	-3334.956	3370.626

```
data frame with 0 columns and 0 rows
```

La chance de que un cliente comita fraude dado que tiene otra tarjeta en el mismo banco es un 68% menor que un cliente que no tiene, cuando el resto de las variables permanecen constantes.

Del mismo modo, un cliente que proporciono un telefono fijo válido tiene una chance de cometer fraude un 71% menor que uno que proporciono uno no válido, cuando el resto de las variables permanecen constantes.

A medida que el crédito límite propuesto aumenta en mil unidades monetarias, la chance de que un cliente cometa fraude aumenta un 191%, manteniendo el resto de variables fijas. VARIFICAR SI ES LINEAL O NO

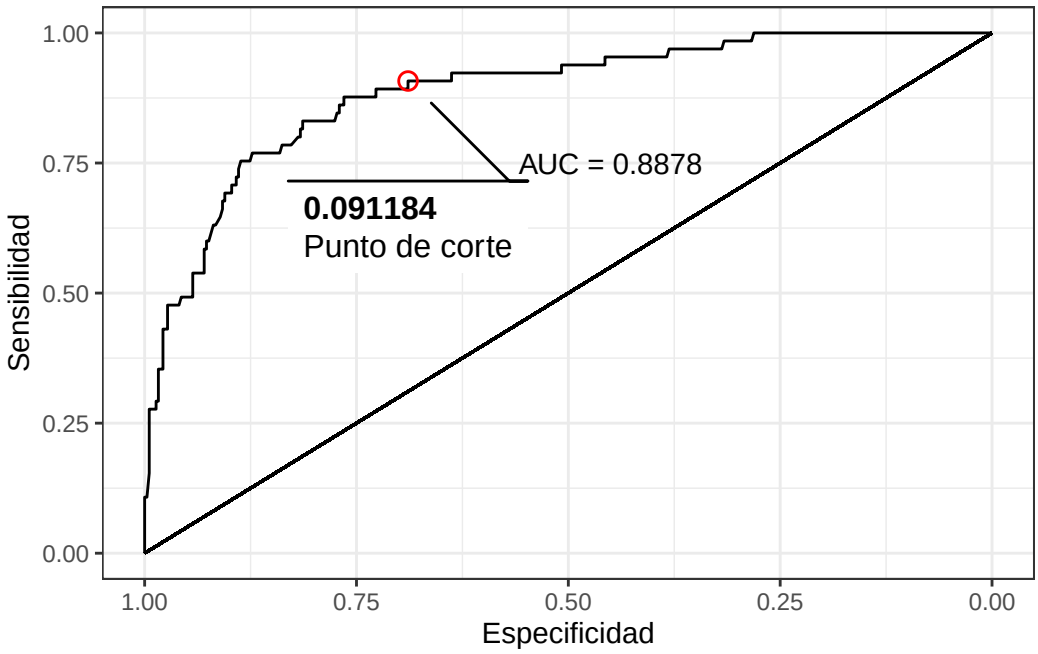
Las características de las personas más propensas a cometer fraude son:

- 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 “BD”
- 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 OSx11
- Un solo email registado en la página del banco en las últimas 8 semanas

Capacidad predictiva



Se decidió utilizar el mejor punto de corte que maximice 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.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 131  29
         1  67  77
```

```
               Accuracy : 0.6842
                 95% CI : (0.6287, 0.7361)
    No Information Rate : 0.6513
    P-Value [Acc > NIR] : 0.1260223

                  Kappa : 0.3582

 Mcnemar's Test P-Value : 0.0001592

            Sensitivity : 0.7264
            Specificity : 0.6616
         Pos Pred Value : 0.5347
         Neg Pred Value : 0.8188
             Prevalence : 0.3487
         Detection Rate : 0.2533
   Detection Prevalence : 0.4737
      Balanced Accuracy : 0.6940

       'Positive' Class : 1


Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 123  22
         1  75  84

               Accuracy : 0.6809
                 95% CI : (0.6253, 0.733)
    No Information Rate : 0.6513
    P-Value [Acc > NIR] : 0.1531

                  Kappa : 0.3706

 Mcnemar's Test P-Value : 1.293e-07

            Sensitivity : 0.7925
            Specificity : 0.6212
         Pos Pred Value : 0.5283
         Neg Pred Value : 0.8483
             Prevalence : 0.3487
         Detection Rate : 0.2763
   Detection Prevalence : 0.5230
      Balanced Accuracy : 0.7068

       'Positive' Class : 1
```