

Modelo logístico con efectos mixtos: aplicación a la detección de fraudes en seguros automotores

Alumno: Ian Bounos

Directora: Dra. Marina Valdora

Co-Directora: Dra. Daniela Rodríguez



Universidad de Buenos Aires

Esquema de presentación

- 1 Conceptos básicos de seguros y motivación del trabajo
- 2 Modelo Logístico con efectos mixtos
- 3 Aplicación práctica a detección de fraudes
- 4 Conclusiones

1 Conceptos básicos de seguros y motivación del trabajo

2 Modelo Logístico con efectos mixtos

3 Aplicación práctica a detección de fraudes

4 Conclusiones

Un contrato de seguro es un acuerdo bilateral en el que una parte, el **asegurador**, se compromete a proteger al otro, el **asegurado**, de ciertos riesgos o pérdidas a cambio de una contraprestación económica periódica denominada **prima**.

Prima pura: el costo esperado de la cobertura del seguro sin tener en cuenta otros factores como los gastos administrativos, impuestos, comisiones a intermediarios y remuneración a la inversión de la empresa.

¿Por qué utilizamos modelos
lineales generalizados (GLM)?
¿Por qué efectos mixtos?

¿Por qué GLM(M)?

En una encuesta mundial realizada por *Akur8*, una empresa multinacional dedicada a brindar servicios de ciencias de datos a empresas de Seguros, **el 90 % de los encuestados afirman que para sus modelos de pricing (cálculos de primas) utilizan GLM's.**

El principal argumento dado por los encuestados para usar este método en lugar de otras técnicas como *Random Forest* o *Gradient Boosting* es su interpretabilidad y facilidad de comunicación de qué es lo que efectivamente realiza el modelo.

¿Por qué es importante la
detección de fraudes en una
empresa de seguros?

Primer motivo: para evitar pérdidas de la empresa aseguradora.

Segundo motivo: hay una segunda razón más sistémica que involucra a todos los agentes participantes: La presencia de fraudes indetectables aumenta de manera artificial el costo esperado del contrato para la aseguradora. Esto generalmente se traslada a una mayor prima pura.

1 Conceptos básicos de seguros y motivación del trabajo

2 Modelo Logístico con efectos mixtos

3 Aplicación práctica a detección de fraudes

4 Conclusiones

Ejemplo ilustrativo: proporción de votantes Obama

Estado	n	y	Proporción
AK	5	3	0.60
AL	29	9	0.31
AR	17	2	0.11
AZ	35	13	0.37
CA	207	129	0.62
CO	37	16	0.43
CT	25	14	0.56
DC	4	4	1.00
DE	6	4	0.66
FL	128	73	0.57

- n_i : cantidad de personas encuestadas en el estado i .
- y_i : cantidad de personas que votarán a Obama en las siguientes elecciones en el estado i

Si solo se dispone de esta información, una primera aproximación es modelar el problema por medio de un **modelo lineal generalizado (GLM)** con distribución binomial y una función link logit, o, en otras palabras, un **modelo logístico**. En términos de ecuaciones puede describirse como:

$$\sum_{j=1}^{n_i} y_{ij} = y_i \sim \text{Bin}(n_i, \pi_i),$$

$$\pi_i = \frac{e^{\beta_i}}{1 + e^{\beta_i}} = g^{-1}(\beta_i),$$

donde:

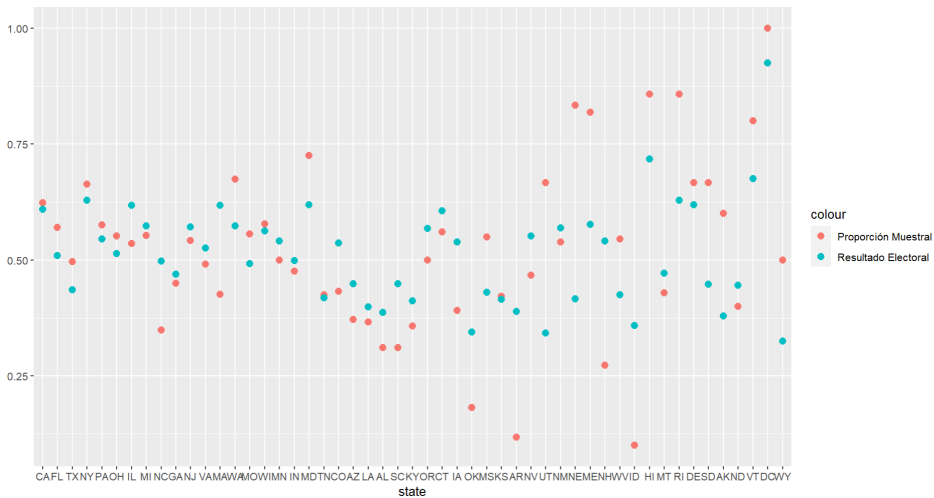
- y_{ij} es el valor observado (es decir, si votará a Obama o no) el individuo j que pertenece al estado i .
- π_i es la probabilidad “real” de que un individuo del estado i vote a Obama
- la función $g(p) = \log\left(\frac{p}{1-p}\right)$, conocida como *logit* es nuestra función *link*.

En este caso, como puede verse en el texto de Agresti¹, el estimador de máxima verosimilitud será la **proporción observada** de los votantes de Obama en cada estado. Es decir:

$$\hat{\pi}_i = \frac{y_i}{n_i}$$

¿Hay un problema?

¹AGRESTI, A. *Categorical Data Analysis*. 3rd Edition. Wiley Series in probability and statistics. 2013.



Definición del modelo

Dados:

- Vectores de covariables explicativas $X_{it} \in \mathbb{R}^p$
- Un vector desconocido de parámetros de efectos fijos correspondientes a las covariables $\beta \in \mathbb{R}^p$
- Un vector de efectos aleatorios por cada clúster $u_i \sim \mathbf{N}_q(0, \Sigma)$, es decir, un vector normal multivariado.
- Vectores observados $Z_{it} \in \mathbb{R}^q$.
- Una función *link* g

Binomial efectos fijos

$$y_{it} \sim \text{Be}(\pi_{it}),$$

$$g(\pi_{it}) = X_{it}^T \beta$$

Binomial efectos mixtos

$$y_{it}|u_i \sim \text{Be}(\pi_{it}),$$

$$g(\pi_{it}) = X_{it}^T \beta + Z_{it}^T u_i$$

Definición del modelo

$$\begin{aligned} P_{\beta, \Sigma}(y_{it} = 1) &= E[y_{it}] = E[E[y_{it}|u_i]] \\ &= E[g^{-1}(X_{it}^T \beta + Z_{it}^T u_i)] = \int g^{-1}(X_{it}^T \beta + Z_{it}^T u_i) f(u_i, \Sigma) du_i \end{aligned}$$

Aquí f es la densidad de una variable normal.

- Si la función link es la identidad se puede probar que $E[y_{it}] = X_{it}^T \beta$. Esto es un argumento para usar link *logit*.
- La integral involucrada suele ser aproximada por métodos numéricos, como la cuadratura de Gauss-Hermite adaptativa ²
- Esto nos permite despejar $P_{\beta, \Sigma}(y_{it} = 0) = 1 - P_{\beta, \Sigma}(y_{it} = 1)$
- También podemos obtener la función de verosimilitud

$$\mathcal{L}_{\beta, \Sigma}(\tilde{y}) = \prod_{it} P_{\beta, \Sigma}(y_{it} = \tilde{y}_{it})$$

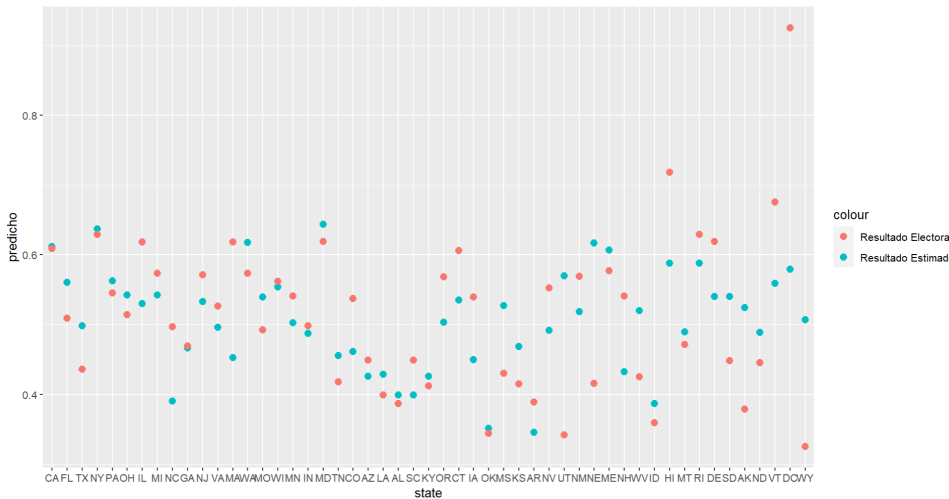
²MCCULLOCH, C. *An Introduction to Generalized Linear Models*. Biometrics Unit and Statistics Center Cornell University. 1997.

Retomando ejemplo ilustrativo

Ilustremos cómo los efectos aleatorios pueden ayudarnos a modelar el ejemplo de las elecciones de 2008. Uno podría considerar el siguiente modelo:

$$\text{logit}(\pi_{it}) = \beta_0 + u_i.$$

En este caso $Z_{it} = X_{it} = 1$ y u_i es una normal univariada de media 0 y varianza σ_u desconocida que como primera aproximación podríamos suponer constante.



Ejemplo 2

Supongamos que en el ejemplo anterior disponemos de la edad de cada votante. Esto nos permite usar la siguiente modificación:

$$\text{logit}(\pi_{it}) = \beta_0 + \beta_1 \text{Edad}_{it} + u_i.$$

En este caso, $X_{it} = \begin{pmatrix} 1 \\ \text{Edad}_{it} \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$ y u_i sigue siendo una perturbación normal propia de cada estado, que, en este caso, no tiene en cuenta la edad.

Ejemplo 3

Lo que podría ocurrir es que en cada estado la edad altere de distinta manera la tendencia a votar a Obama. Eso se puede modelar de la siguiente forma:

$$\text{logit}(\pi_{it}) = \beta_0 + \beta_1 \text{Edad}_{it} + u_{1i} \text{Edad}_{it} + u_{0i}$$





Aquí, X_{it} y β son exactamente los mismos que en el anterior, pero $Z_{it} = \begin{pmatrix} 1 \\ \text{Edad}_{it} \end{pmatrix}$ y $u_i = \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix}$

¿Efectos fijos o aleatorios?

- ① Si una variable es controlada o manipulada por el investigador, entonces es un motivo para considerarla un efecto fijo. Por ejemplo, en un estudio sobre el efecto de una droga en la supervivencia de pacientes con cáncer, la dosis de la droga sería un efecto fijo.
- ② Si la variable es un factor con un número limitado de niveles, entonces se considera un efecto fijo. Un posible ejemplo sería el sexo de un animal en un estudio de laboratorio. En cambio, si la variable es un factor con un número ilimitado de niveles entonces se lo puede considerar un efecto aleatorio. Ejemplo: la marca de un automóvil.

¿Efectos fijos o aleatorios?

- 1 Es preciso tener en cuenta la naturaleza de la variable: Si esta es una característica que se asigna a un grupo o individuo, como la edad o el género, es probable que sea un efecto fijo. Por otro lado, si la variable es una medida tomada en diferentes individuos o grupos, como el peso o la altura, se podría incorporar un efecto aleatorio en el cual la pendiente que multiplica a la medida en cuestión dependerá del individuo o grupo en el cual se está efectuando la medición.
- 2 Cuando los datos están desbalanceados y se cuenta con pocas observaciones de un determinado grupo o individuo, puede ser recomendable utilizar efectos aleatorios porque, como se vio anteriormente, realiza un *trade off* entre la información dada por el grupo y la dada por la totalidad de las observaciones teniendo en cuenta la cantidad de ejemplares de cada uno. Sin embargo, cabe resaltar que esto, en algunas condiciones puede llevar a sesgos ³

³GEORGE, ROSENBAUM, ET AL. *Mortality Rate Estimation and Standardization for Public Reporting: Medicare's Hospital Compare*. University of Pennsylvania. 2016.    

Up with BLUP

*No ma'am, No ma'am. No one
knows ma'am. Whether FIXED
or whether RANDOM.*

*Should we choose this one, or do that one?
I.I.D. says that they're RANDOM.*

*But if we must say that's nixed,
then we'll say that they are FIXED.*

*What of those who won't predict?
To them I say interdict.*

*Up with BLUP,
BLUP's for you.*

That's for when you're interested too.⁴

⁴MCCULLOCH, C. *An Introduction to Generalized Linear Models*. Biometrics Unit and Statistics Center Cornell University. 1997.

1 Conceptos básicos de seguros y motivación del trabajo

2 Modelo Logístico con efectos mixtos

3 Aplicación práctica a detección de fraudes

4 Conclusiones

Disponemos de un dataset con 1000 observaciones de 40 variables relativas siniestros de automóviles ocurridos en Estados Unidos. El mismo está públicamente disponible y fue obtenido en Kaggle⁵

⁵El dataset puede encontrarse en <https://www.kaggle.com/code/bunttyshah/insurance-fraud-claims-detection>

- **Datos del asegurado:**

- Edad
- Género (M o F)
- Nivel educativo, profesión, hobbies, estado civil
- Meses como asegurado

- **Datos de la póliza:**

- Fecha de inicio de vigencia de la póliza
- Estado correspondiente (solo tenemos datos de Ohio, Illinois e Indiana)
- Valor del deducible
- Valor de la prima

- **Datos del auto:**

- Marca
- Modelo
- Año

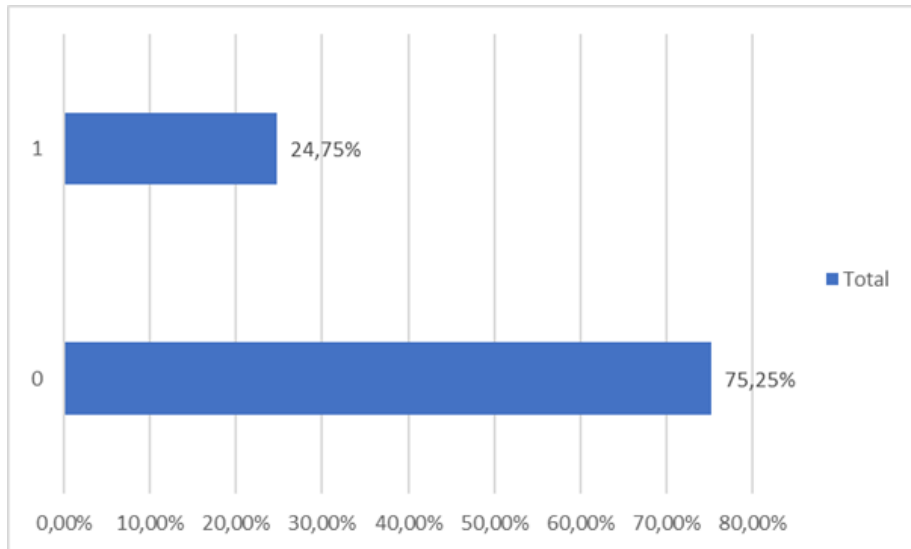
Datos del Siniestro:

- Fecha de ocurrencia
- Monto del siniestro en conceptos de daños por lesiones a terceros y reparaciones patrimoniales (*Property* y *Vehicle*).
- Tipo de incidente (Robo, colisión, etc.)
- Severidad del accidente (Daños menores, mayores, totales, triviales)
- Autoridades contactadas (Policía, Ambulancia, etc.)
- Ubicación del incidente (Estado, ciudad y dirección)
- Hora del incidente.
- Cantidad de Vehículos involucrados
- Cantidad de cuerpos humanos dañados
- Cantidad de testigos
- Variable dummy que indica si hubo daños a propiedades no automotores y otra que indica el monto.
- Variable dummy que indica si hubo reporte policial

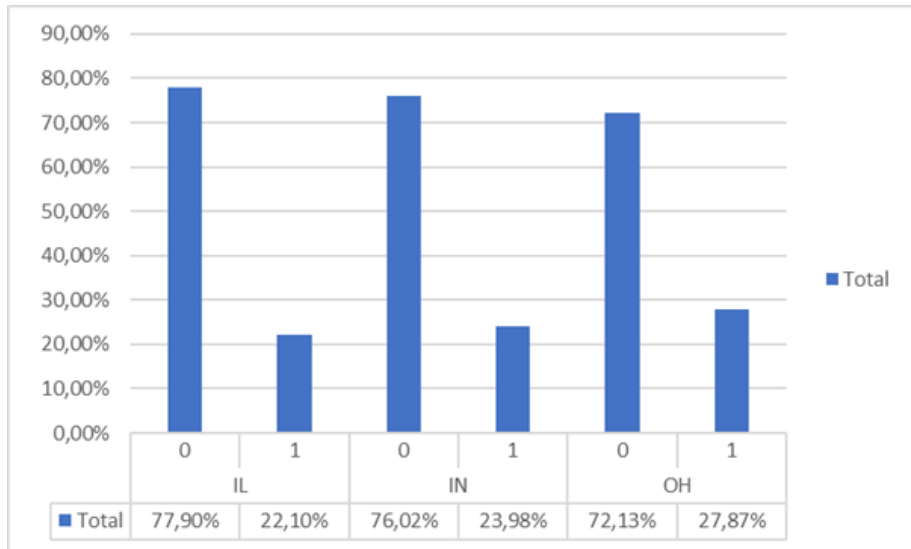
“Fraud Reported” Esta es nuestra variable dependiente a explicar. Toma el valor 1 si se detectó Fraude y 0 en el caso contrario.

Análisis exploratorio para la preselección de variables

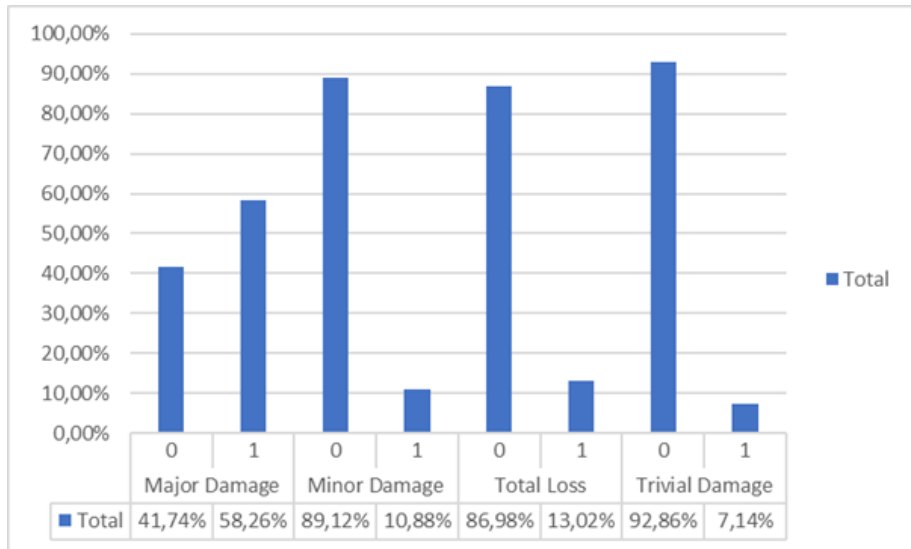
Análisis Exploratorio



Análisis Exploratorio: Variables cualitativas



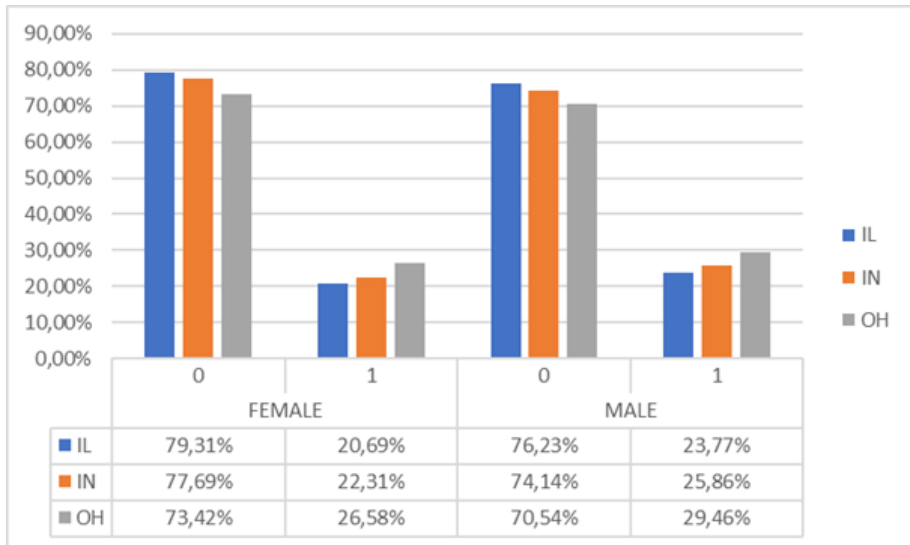
Analisis Exploratorio: Variables Cualitativas



Variables preseleccionadas:

- Severidad del Incidente
- Tipo de incidente
- Tipo de Colisión
- Autoridades contactadas
- Estado en el que ocurrió el accidente (el cual no necesariamente es el mismo que el de la póliza)
- Cantidad de Testigos
- Marca del Automóvil
- Hobbies del Asegurado

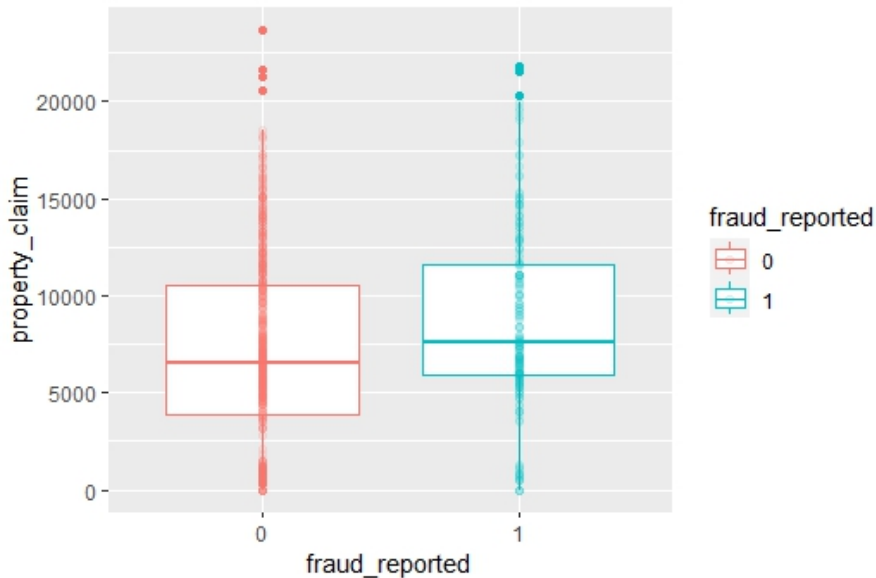
¿Interacciones?



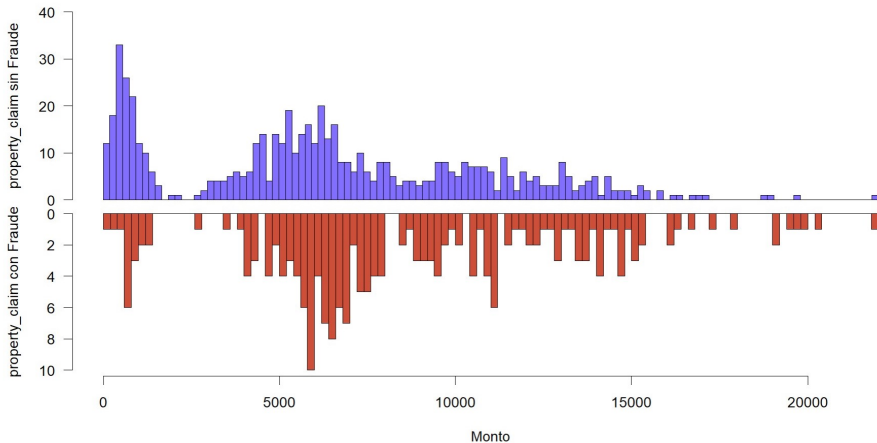
Para el análisis de las variables cuantitativas se ha utilizado la herramienta del diagrama de cajas o boxplot. En términos generales, no se han encontrado variables cuya influencia se perciba gráficamente, con la excepción de:

- Monto de daños físicos a terceros (*injury_claims*)
- Monto de daños patrimoniales no relacionados al vehículo (*property_claims*)
- Monto de daños patrimoniales relacionados con el vehículo (*vehicle_claims*)

Análisis Exploratorio: Variables Cuantitativas



Análisis Exploratorio: Variables Cuantitativas



	No Fraude	Fraude
Property Barato	135 (89 %)	17 (11 %)
Property Caro	467 (72 %)	181 (28 %)

Cuadro: Tabla de frecuencias absolutas Property-Fraude.

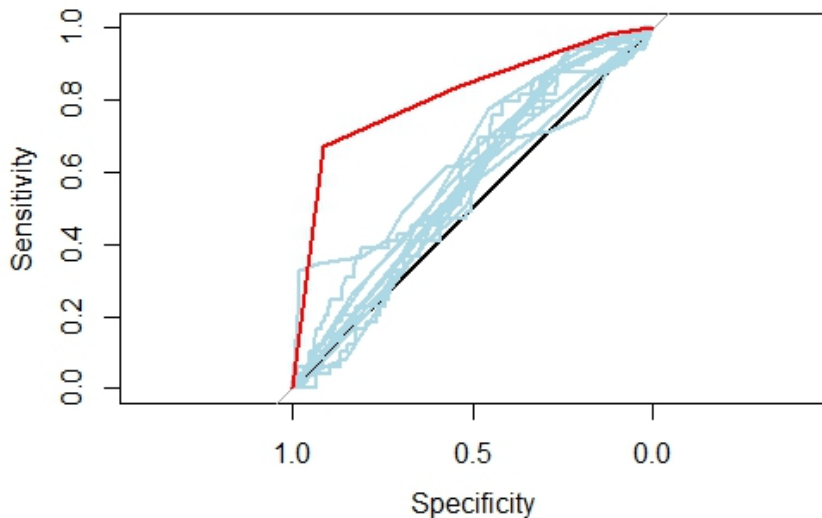
Esto sugiere incorporar la variable dummy *property_caro*. El mismo fenómeno ocurre con los demás tipos de siniestros.

Selección de variables

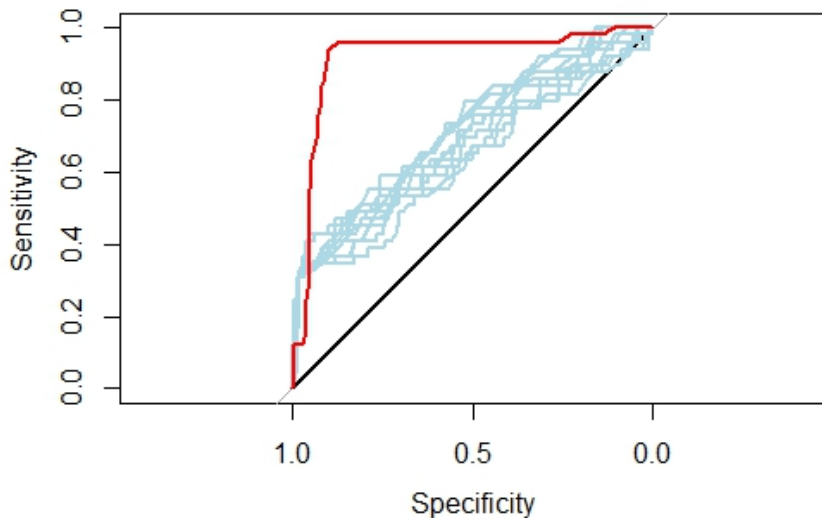
Selección de variables con el criterio de maximización de AUC con un método Forward

- ➊ **Inicialización:** Separamos a nuestro conjunto en un subconjunto de entrenamiento y testeo. Le calculamos el AUC al modelo que no incluye ninguna de las covariables fuera del intercepto (con lo cual nuestro conjunto de covariables inicial es nulo). Este será nuestro AUC base.
- ➋ **Cálculo de AUC:** Para cada uno de los candidatos a variables explicativas lo agregamos provisoriamente al conjunto de covariables y calculamos el AUC, **considerándolo como efecto fijo o aleatorio según se haya predefinido.**
- ➌ **Actualización:** Si el máximo de los AUC calculados en el paso anterior resulta mayor que el AUC base con una diferencia mayor que 0,01, actualizamos el AUC base a dicho valor, agregamos la variable al conjunto de covariables y la quitamos del conjunto de candidatos a variables explicativas.
- ➍ **Criterio de Cierre:** Si el AUC base no se modificó en el paso anterior, el proceso termina; si no, se vuelve al paso 2.

Selección AUC: Paso 1



Selección AUC: Paso 2



Selección por criterio AUC

En el primer paso, seleccionamos la variable *incident_severity* como efecto fijo con un AUC de 0.81. En el segundo paso, se elige *insured_hobbies* como efecto aleatorio, lo cual sube el AUC a 0.91. El proceso termina ahí y define el siguiente modelo:

$$\begin{aligned} \text{Fraude}_{it} | u_i &\sim \text{Be}(\pi_{it}) \\ \text{logit}(\pi_{it}) &= \beta_0 + \beta_1 \text{Severidad}_{it} + u_i \\ u_i &\sim \mathcal{N}(0, \sigma_u^2) \quad \text{iid} \end{aligned}$$

Selección de variables: Enfoque por AIC

Otro enfoque posible de selección de variables sería el de minimización de Criterio de Información de Akaike (AIC). Al igual que en AUC, la dirección del proceso será *forward*.

Paso	Variable incorporada	AIC	Var. AIC	Deviance	P-val test Chisq
0	(Intercept)	897.31	-	895.32	-
1	incident_severity	718.94	178.36	710.95	<2e-16 ***
2	insured_hobbies	625.75	93.19	615.75	<2e-16 ***
3	collision_type	625.57	0.18	609.58	0.10

Cuadro: Resultados del proceso de selección forward por minimización AIC

Con respecto a “tipo de colisión”: por su bajo aporte al AIC, su no significatividad del test de deviance y por evitar complejizar innecesariamente el modelo será excluida del análisis y el modelo propuesto será el ya mencionado, que maximiza el AUC.

Análisis de Resultados

Estimación y significatividad individual

Efectos Fijos	Estimación	Std. Err.	P valor
(Intercept)	0.5612	0.3171	0.0767.
incident_severity Minor Damage	-3.047	0.2858	< 2e-16 ***
incident_severity Total Loss	-2.9827	0.3065	< 2e-16 ***
incident_severity Trivial Damage	-3.5058	0.5505	1.91e-10 ***

Cuadro: Estimación de Efectos Fijos para el Modelo

Significatividad global

Modelo	Deviance	Dif Deviance	P-val test
Modelo sin incident_severity	832.98	-	-
Modelo Propuesto	615.75	217.23	<2e-16 ***

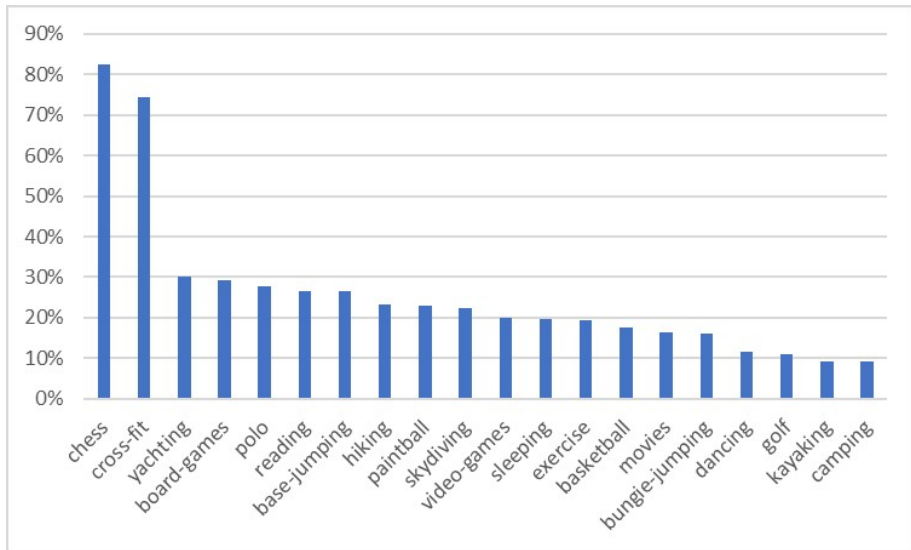
Cuadro: Test cociente de verosimilitud para *incident_severity*

Predictores de hobbies

Hobby	Pred. Modelo
chess	3.2679382
cross-fit	2.6841093
camping	-1.7190576
kayaking	-1.1210522
sleeping	-0.8395051
exercise	-0.6484961
golf	-0.6395591
paintball	-0.6056873
bungee-jumping	-0.5931328
yachting	0.5789181
dancing	-0.4969491
hiking	0.4918120

Cuadro: Predicción de Efectos Aleatorios

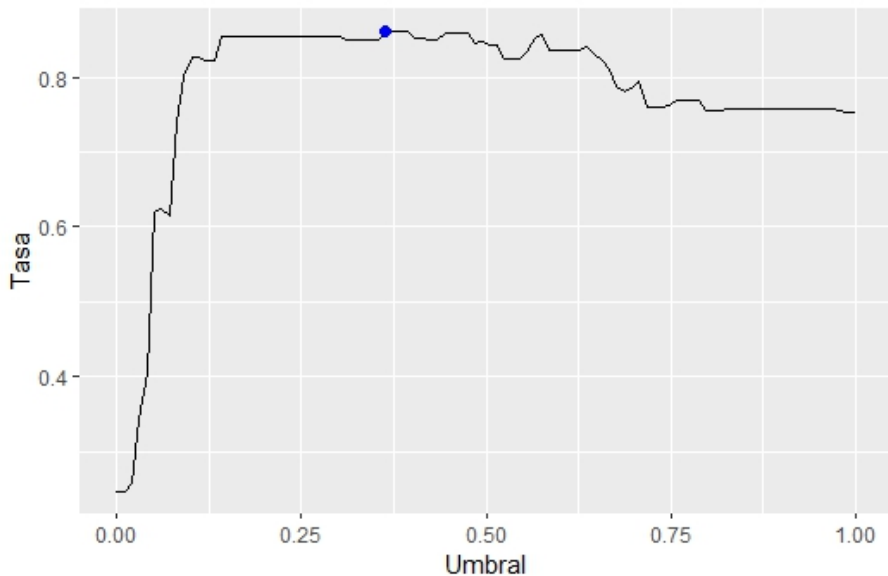
Predictores de hobbies



La primera métrica a evaluar será la **tasa de aciertos**. Todas las métricas se calculan con LOOCV.

Si en cada modelo se clasificara una observación como fraudulenta, si la probabilidad estimada es mayor que el 50 %, la tasa de aciertos sería 84,8 %. Sin embargo, por muchos motivos podríamos optar por cambiar este umbral de probabilidad. Por ello, graficamos en cuál es la tasa de aciertos del modelo en función del umbral de probabilidad. Este óptimo empírico se da cuando el umbral de probabilidad es 36 % y la tasa de aciertos resulta 86,1 %.

Tasa de aciertos



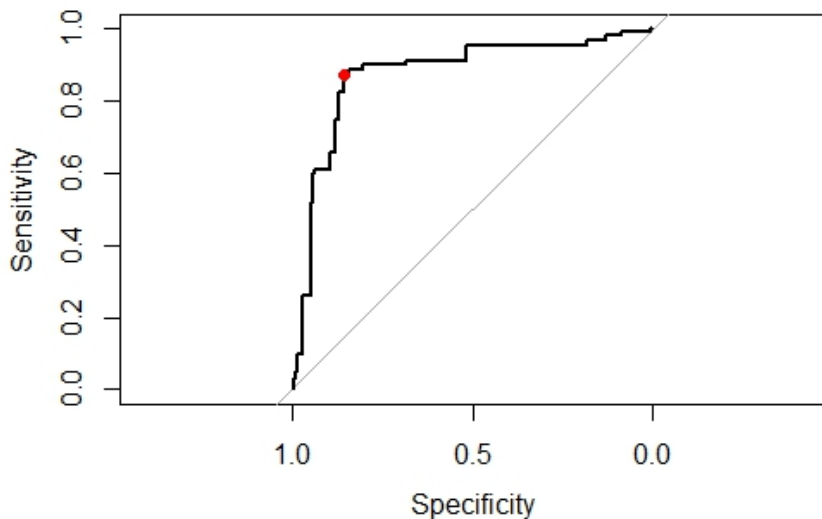
Matriz de confusión

	No Fraude	Fraude
Predicho No Fraude	646	32
Predicho Fraude	107	215

Cuadro: Tabla de Confusión del Modelo con el umbral de probabilidad que maximiza la tasa de aciertos.

Esta tabla lleva implícita una sensibilidad del 87,0 % y una especificidad de 85,7 %

Curva ROC con LOOCV



1 Conceptos básicos de seguros y motivación del trabajo

2 Modelo Logístico con efectos mixtos

3 Aplicación práctica a detección de fraudes

4 Conclusiones

Hemos:

- 1 Realizado un análisis exploratorio para:
 - 1 preseleccionar variables
 - 2 generar nuevas variables
- 2 Seleccionado variables con base en la preselección anterior con maximización de AUC y criterio Forward.
- 3 Definido un modelo logístico con efectos mixtos que tiene la ventaja de poder realizar predicciones de asegurados con hobbies fuera de la base de datos.
- 4 Obtenido una tasa de aciertos del 86.1 % haciendo uso de LOOCV, con lo cual no hay riesgo de sobreajuste.