Predicción de incumplimientos crediticios en PYMEs

Regresión Avanzada

David Edgardo Castillo Rodríguez
Miguel Ángel Ávila del Bosque
Jorge III Altamirano Astorga
Mario Alberto Cruz García

ITAM

Problema

Aunque son muy importantes las PYMEs, pues según la CONDUSEF:

- Aportan al **72%** de los empleos en el país
- Aportan **52%** al PIB

Aún así no se cuentan con mecanismos de financiamiento, debido a la escasa información financiera con la que existe.

Surge la necesidad de poder conocer los riesgos y liquidez de dicho sector empresarial.

ITAM

Propuesta

Implementar soluciones **predictivas** que permitan conocer el conjunto de variables que puedan describir y predecir.

Con la finalidad de poder definir y gestionar **riesgos** probables, definir políticas y establecer acciones que permitan la recuperación efectiva de los fondos invertidos en cuentas por cobrar.

Llamaremos el **riesgo de crédito** como:

> “La probabilidad de que, al vencimiento del crédito, el cliente no cumpla (default) en su totalidad o parcialmente sus compromisos u obligaciones contraídos por falta de liquidez.”

ITAM

Datos: Separación en 2 conjuntos

Separamos nuestros datos con una semilla estática, para hacerlo reproducible:

1. Observaciones utilizadas para el conjunto de entrenamiento (~70%): 10595
2. Observaciones utilizadas para el conjunto de para prueba (~30%): 4542

Realizamos muestras con un promedio en la variable respuesta como se muestra a continuación al dividir el conjunto de datos original.

1. Media de la "y" en el conjunto original: 0.1936
2. Media de la "y" en el conjunto de entrenamiento: 0.1932
3. Media de la "y" en el conjunto de pruebas 0.194

ITAM

Datos: EDA

Gráfico de frecuencia absoluta variable bkicq

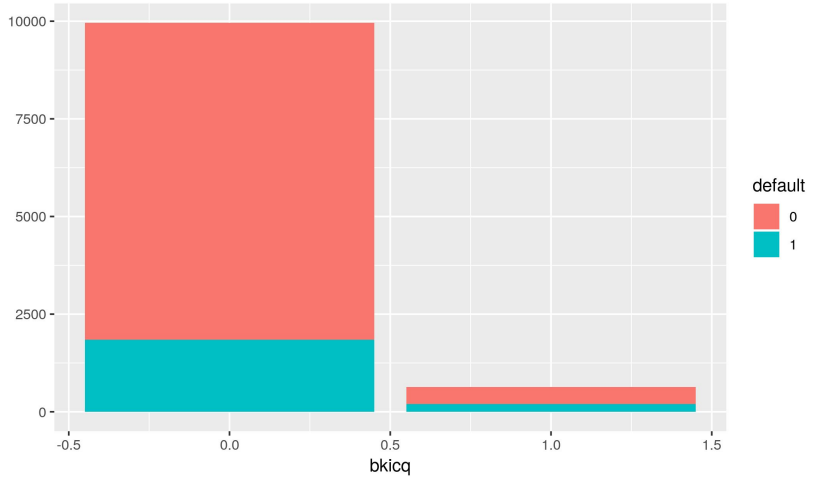
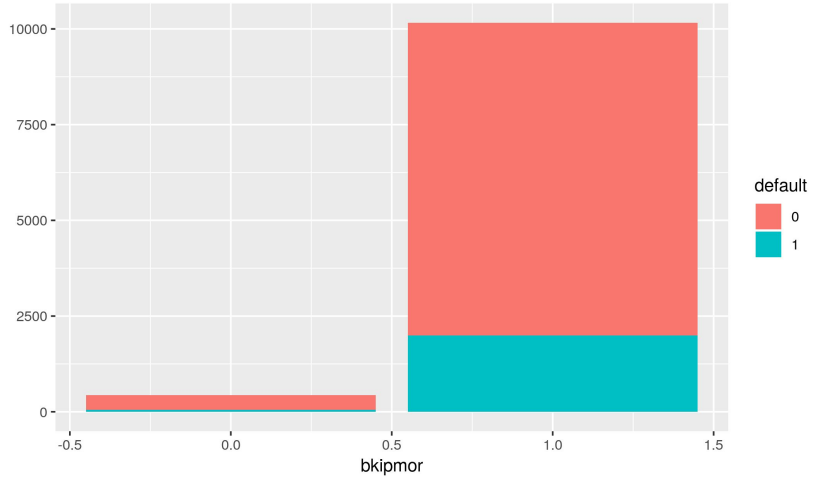


Gráfico de frecuencia absoluta variable bkipmor



Datos: EDA

Gráfico de frecuencia absoluta variable nbkpromt

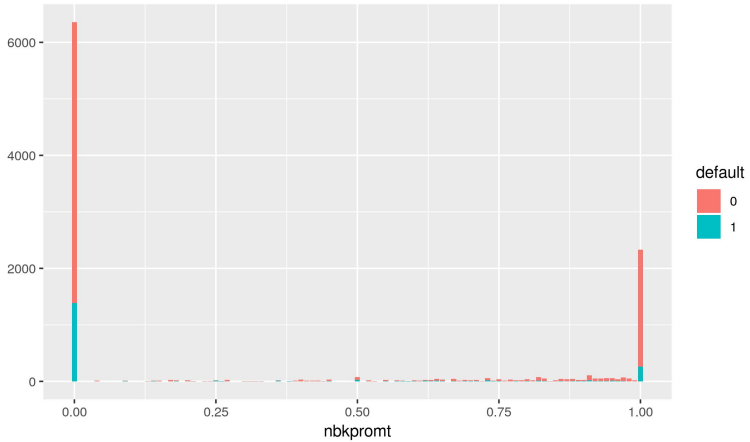
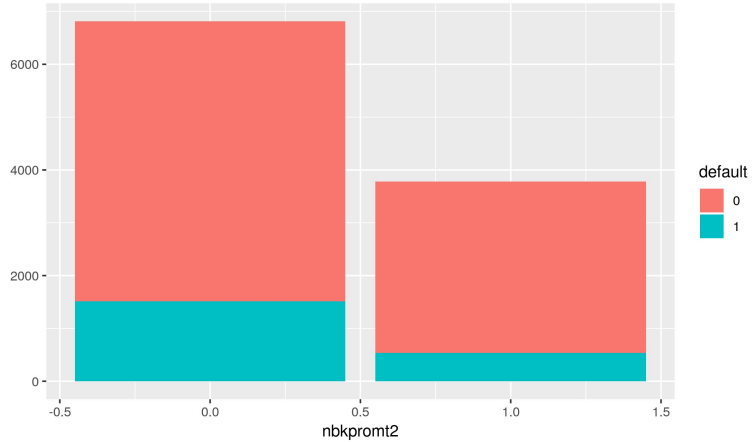


Gráfico frecuencia absoluta variable nbkpromt2



Datos: EDA

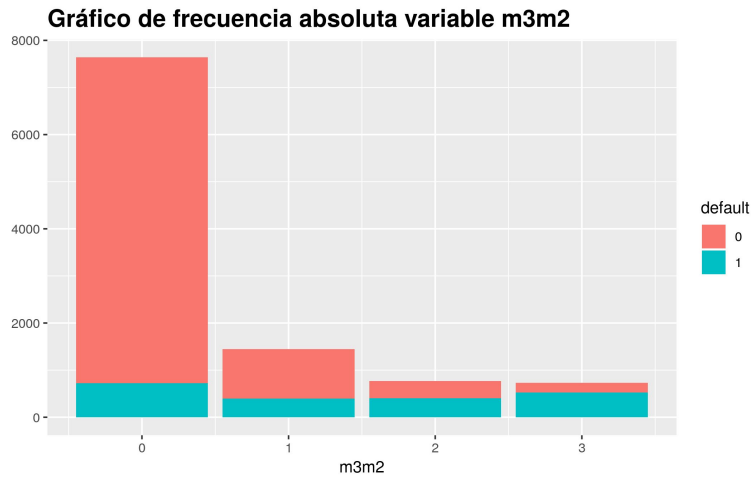


Gráfico de frecuencia absoluta variable m3m2
8000
6000
4000
2000
0
0
1
2
3
m3m2
default
0
1

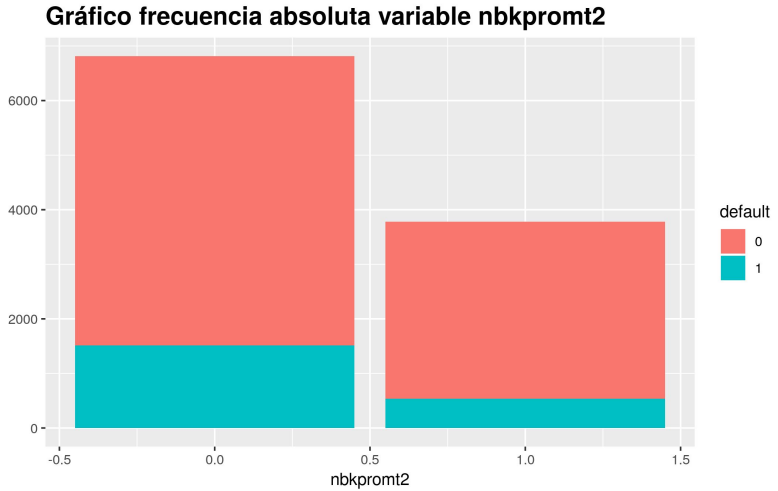


Gráfico frecuencia absoluta variable nbkpromt2
6000
4000
2000
0
-0.5
0.0
0.5
1.0
1.5
nbkpromt2
default
0
1

Datos: EDA

Gráfico de frecuencia absoluta variable bkpromt

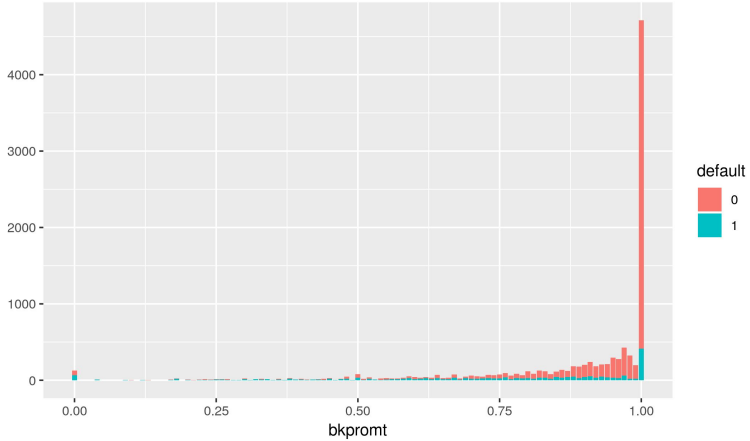
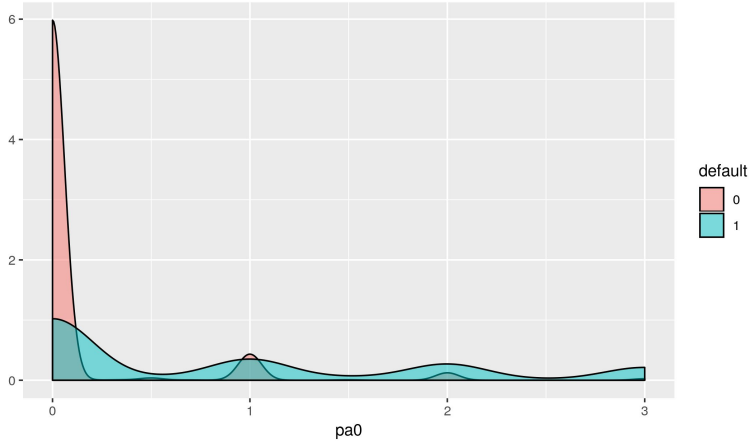
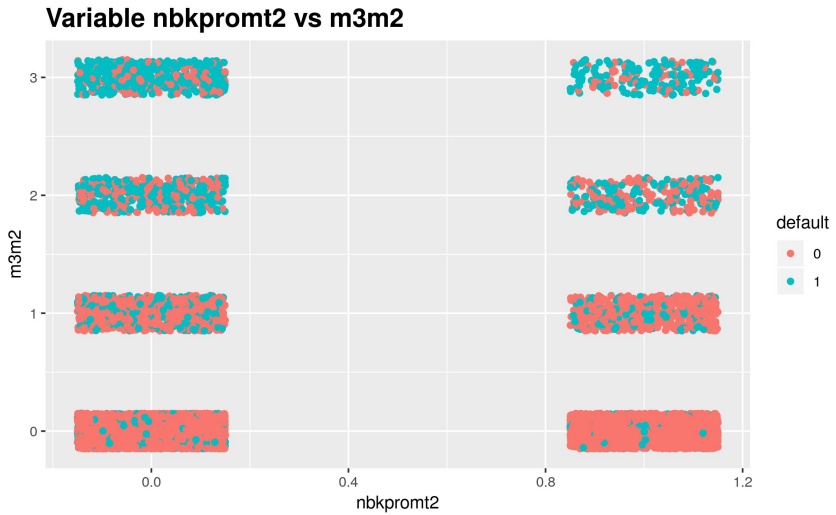


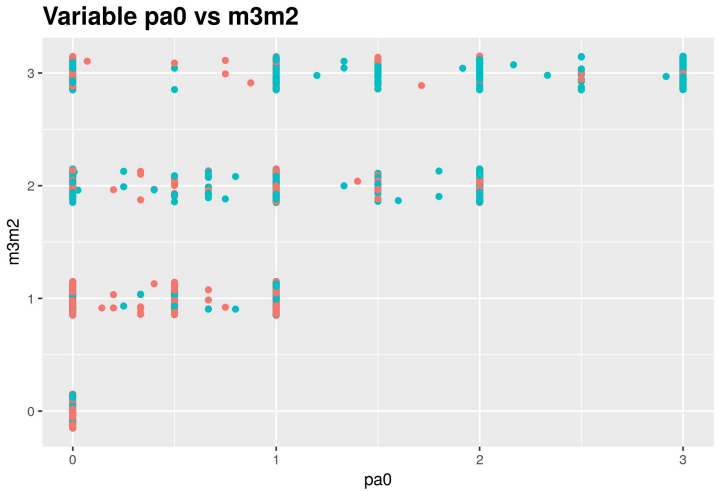
Gráfico de densidad variable pa0



Datos: EDA


Variable nbkpromt2 vs m3m2
m3m2
3
2
1
0
0.0
0.4
0.8
1.2
nbkpromt2
default
0
1

Datos: EDA


Variable pa0 vs m3m2
m3m2
pa0
default
0
1

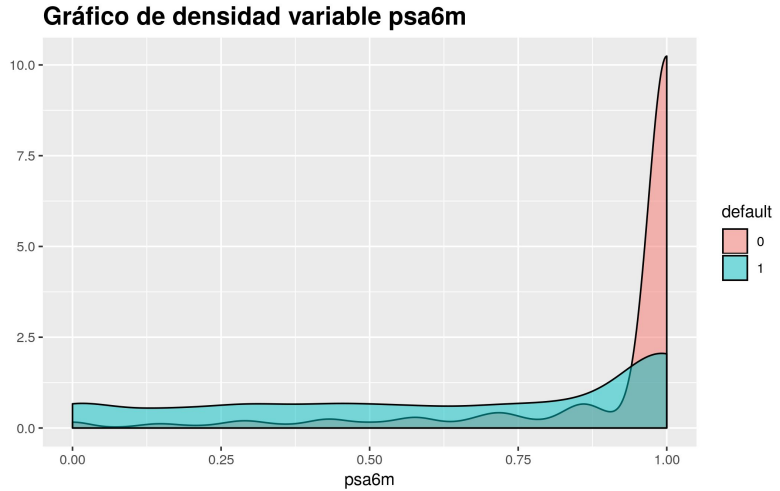


Gráfico de densidad variable psa6m
psa6m
default
0
1

Datos: PCA

Grupo 1	Grupo 2	Grupo 3	Grupo 4
pa0	m3m	mofb	bkpromt
	bkicq	nbkpromt	psa6m
		bkipmor	

Datos: Variables del Modelo

Variable	Descripción
y	Marca de incumplimiento (vale 1 si el crédito fue declarado en default y 0 e.o.c.)
m3m2	Máximo número de atrasos en los 3 meses anteriores
nbk_promt2	% pagos en tiempo en los últimos 12 meses a instituciones financieras no bancarias
bkpromt	% pagos en tiempo en los últimos 12 meses a instituciones financieras bancarias
pa0	Promedio de atrasos a tiempo 0
nbm3	Creada mediante ingeniería de variables: $(nbkpromt2 + .01) * ifelse(\text{m3m} \geq 3{,}3\ , \text{m3m})$

Modelos

Se probaron modelos con verosimilitud Bernoulli y con cada una de las funciones liga mostradas en clase con el fin de comparar los mejores DICs.

$$y_i \mid \mu_i \sim Bernoulli(\theta)$$

$$\eta = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4 + \beta_6 X_5$$

Liga Logística: $\theta = \frac{1}{1+e^{\eta}}.$

Liga C-Log Log: $\theta = log(-log(\eta)).$

Liga Probit: $\theta = \Phi(\eta).$

Liga Log Log: $\theta = log(-log(1-\eta)).$

ITAM

Modelos

$$\beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4 + \beta_6 X_5$$

X_1: m3m2

X_2: nbkpromt2

X_3: bkpromt

X_4: pa0

X_5: nb3

Comparación de Modelos

Aquí se pueden observar el desempeño de los mencionados modelos con las medias de los coeficientes lo cual se logró con 20,000 simulaciones cada uno en el muestreador de Gibbs para asegurar convergencia:

Modelo	DIC	β_1	β_2	β_2	β_3	β_4	β_5
LogLog	7958.21	0.56	-0.29	0.11	0.25	-0.28	-0.01
C-LogLog	8134.63	-1.83	0.44	-0.27	-0.21	0.24	0.11
Probit	8341.72	-1.01	0.3	-0.15	-0.22	0.24	0.05
Logit	8494.03	-1.72	0.52	-0.28	-0.38	0.41	0.09

Interpretación de Resultados

Se interpretarán a continuación los coeficientes. Para lo cual tomamos el modelo Bernoulli con liga log log, dado que obtuvo el mejor desempeño basándonos en su DIC, pudiéndose expresar como:

$$log(-log(\mu_i)) = \eta_i = x_i\beta \iff \mu_i = exp\{exp[x_i\beta]\}$$

Tenemos que:

$$\frac{log(\mu_j)}{log(\mu_i)} = e^{\beta_i}$$

Conclusiones

Se cumplió el objetivo de obtener un *score* basado en la **probabilidad de default** con **resultados razonables**.

Sin embargo, podemos decir que **se requieren más pruebas en más créditos** para poder tener mayor certidumbre de que nuestro modelo efectivamente fomentaría que los créditos sean dados a las PYMEs, manteniendo la certeza de que van a ser pagados aún sin un historial crediticio.

También mediante el análisis de los datos se logró el **objetivo** de **identificar** a un **conjunto de variables** que pudieran **describir** y **predecir** las **dificultades financieras** de las empresas.

ITAM

Referencias

- Luis E. Nieto-Barajas. Notas del curso de regresión avanzada. 2019.
- Nicky Best Dave Lunn David Spiegelhalter, Andrew Thomas. OpenBUGS User Manual. Cambdridge University, March 2014.
- Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian data analysis. Chapman and Hall/CRC, 2013.
- David Lunn, Chris Jackson, Nicky Best, David Spiegelhalter, and Andrew Thomas. The BUGS book: A practical introduction to Bayesian analysis. Chapman and Hall/CRC, 2012.

ITAM

Anexos

Convergencia
de las Cadenas