# Predicción de incumplimientos crediticios en PYMEs

David E. Castillo, Miguel A. Ávila, Jorge III Altamirano, Mario A. Cruz

*Instituto Tecnológico Autónomo de México*

Diciembre 2019

## 1. Introducción

### 1.1. Importancia del Risk Ranking

En América Latina, actualmente las pequeñas y medianas empresas constituyen el mayor factor empresarial, según la CONDUSEF generan el 72 % de los empleos del país y el 52 % del producto interno bruto. En México hay más de 4.1 millones de PYMEs. Se concentran en actividades como el comercio, los servicios, la industria artesanal y trabajos independientes.

A pesar de la notable importancia en la economía nacional de estas empresas, no cuentan con un apoyo adecuado que promueva el desarrollo y el crecimiento. Una de estas causas que lo restringen son las limitaciones a las fuentes de financiamiento formales para el desarrollo adecuado de sus operaciones.

Esto debido a la falta o escasa información financiera que existe que permita evaluar la capacidad de ser sujeto de crédito, como también disponer de activos de calidad que garanticen sus obligaciones ni disponer de un historial financiero que permita evaluar la capacidad de pago.

Por lo anterior, surge el interés de conocer conocer el riesgo de crédito que tienen las empresas. Siendo lo anterior uno objetivos del proyecto, pues brindaría de un instrumento adecuado para tomar mejores decisiones y actuar oportunamente.

A pesar del desarrollo de modelos predictivos de la capacidad de pago, un problema común es la identificación del conjunto de variables que puedan describir y predecir, de forma fiable las dificultades financieras de las empresas.

El uso de información financiera, es de gran utilidad, en el análisis y gestión del crédito, para definir políticas de ventas, inversiones, condiciones de pagos, identificar y gestionar el riesgo de improbables y establecer acciones que permitan asegurar la recuperación de los fondos invertidos en cuentas por cobrar.

Debido a esto, se puede tomar la gestión de créditos como una actividad riesgosa y clave en la gestión de fondos dados los efectos en la liquidez y solvencia del negocio.

Se entiende como riesgo de crédito a la probabilidad de que, a su vencimiento un cliente no cumpla, en su totalidad o parcialmente, sus obligaciones o compromisos contraídos debido a la falta de liquidez.

Con la finalidad de disminuir los riesgos crediticios, es decir, la recuperación de las cuentas por cobrar, es recomendable analizar la industria en la cual participa el cliente, el comportamiento histórico y la capacidad potencial de pago.

# 2. Descripción de la información

## 2.1. Selección de variables

Para el presente proyecto se realizó un análisis de las variables para saber cuáles eran las que tenían mayor poder explicativo sobre la variable de interés. En concreto un análisis exploratorio de datos y uno de componentes principales.

## 2.2. Datos

La base de datos consta de un total de 15,139 observaciones y cada observación corresponde a un crédito otorgado a alguna PYME.

Las variables a analizar son las siguientes:

| **Abreviatura** | **Variable** | **Descripción** |
|---|---|---|
| y | incumplido_prueba | Vale 1 si el crédito fue declarado en default y 0 e.o.c. |
| m3m | max_atraso_3m | Máximo número de atrasos en los 3 meses anteriores |
| psa6m | pje_sdo_atr0_6m | Porcentaje de saldo con atraso en los 6 meses anteriores |
| mofb | months_on_file_banking | Meses desde que la empresa fue reportada en buró por primera vez |

| Abreviatura | Variable | Descripción |
|---|---|---|
| nbkpromt | nbk12_pct_promt | Porcentaje de pagos en tiempo en los últimos 12 meses a instituciones financieras no bancarias |
| bkpromt | bk12_pct_promt | Porcentaje de pagos en tiempo en los últimos 12 meses a instituciones financieras bancarias |
| bkicq | bk12_ind_qcra | Indicador de quitas y castigos |
| bkipmor | bk_ind_pmor | Indicador de persona moral |
| pa0 | prom_atr0 | Promedio de atrasos a tiempo 0 |

## 2.3. Separación en Datos de Entrenamiento y Pruebas

Como buena práctica de Ciencia de Datos aplicamos una separación de nuestros datos con una semilla estática, para hacerlo reproducible. Dividimos de la siguiente manera:

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

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

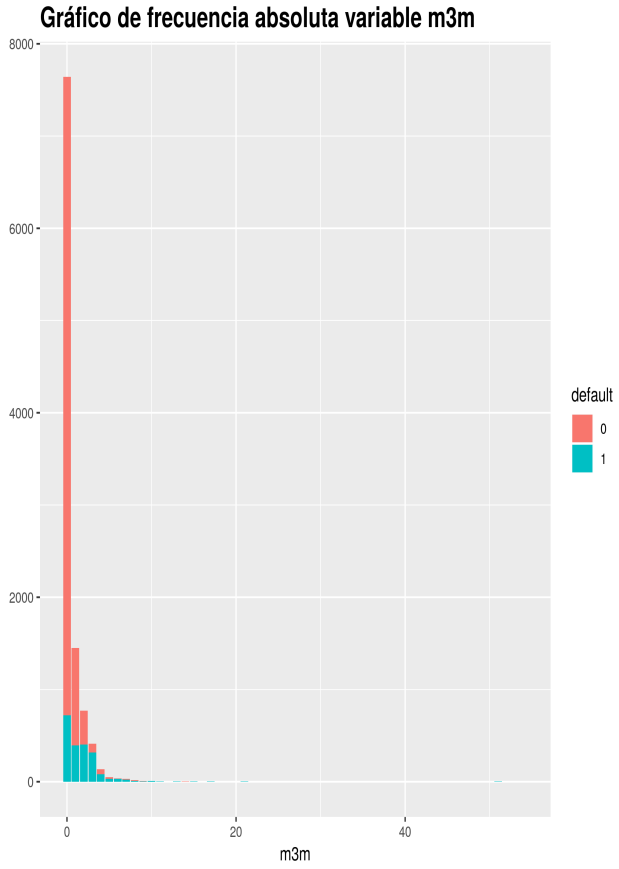
1. Media de la 'y' en el conjunto original: 0.1936
2. Media de la 'y' en el conjunto de entrenamiento: 0.1932
3. Media de la 'y' en el conjunto de pruebas 0.194

Utilizamos el conjunto de entrenamiento para ajustar el modelo.

## 2.4. EDA: Análisis exploratorio de datos

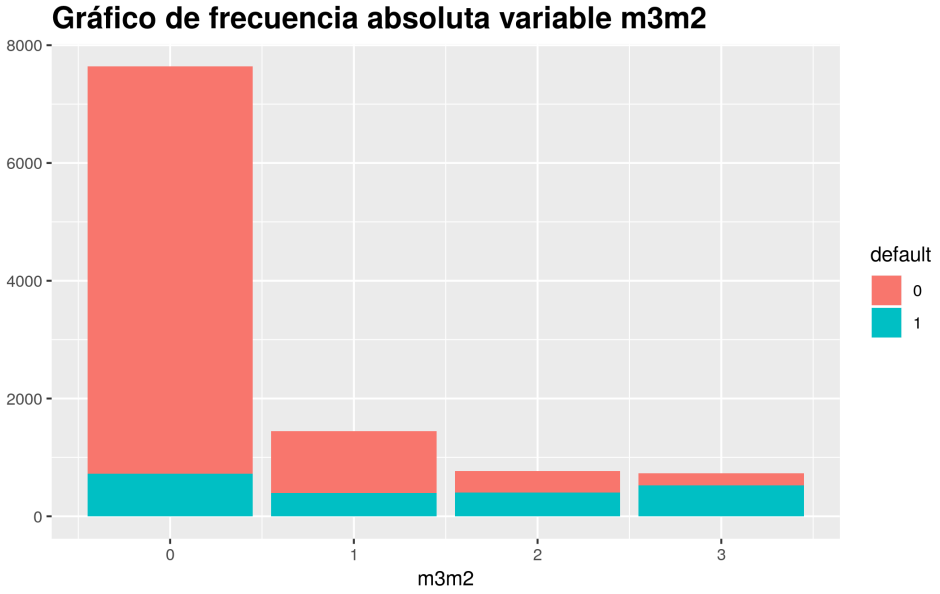
A continuación se muestran gráficos de frecuencia absoluta así como de densidad, dependiendo el tipo de escala de medición de cada variable, con el fin de explorar cómo se comporta el número de incumplimientos (defaults) en el rango de cada variable. Esto nos ayudará a proponer un primer conjunto de variables (así como encontrar interacciones entre estas).

En el siguiente gráfico se puede observar que la variable m3m está fuertemente sesgada hacia la derecha, además pareciera que hay varios valores atípicos, por ello se propone agrupar esta variable, como todas aquellas cuentas que tomen valores mayores o iguales a 3 considerarlas como si su valor fuera igual a 3, esto último puede apreciarse en el gráfico 2. La cual se denominará m3m2.

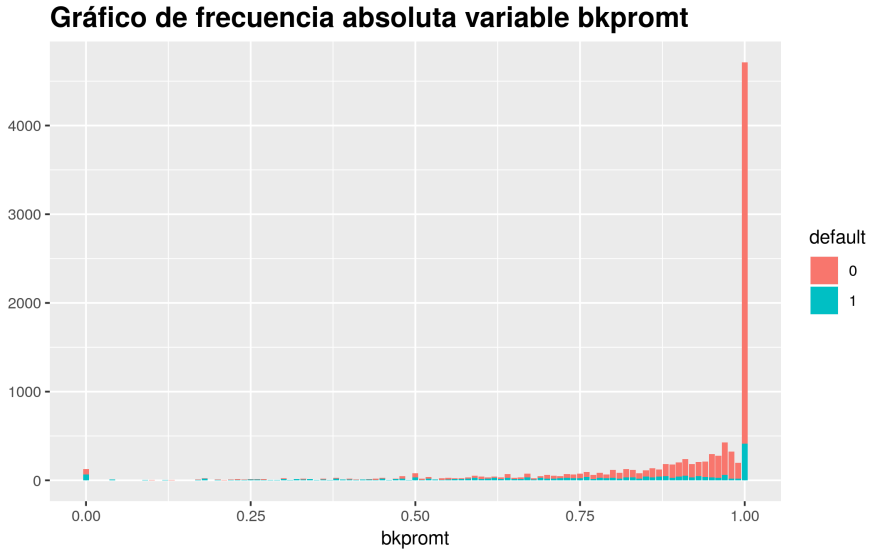



En el siguiente gráfico se muestra la variable m3m2, la cual se construyó como se especifica en el párrafo anterior. Es importante mencionar que cuando esta variable toma el valor de 0, se tiene una tasa de default del 10 %, cuando toma el valor de 1 la tasa de default es del 25 %, cuando toma el valor de 2 la

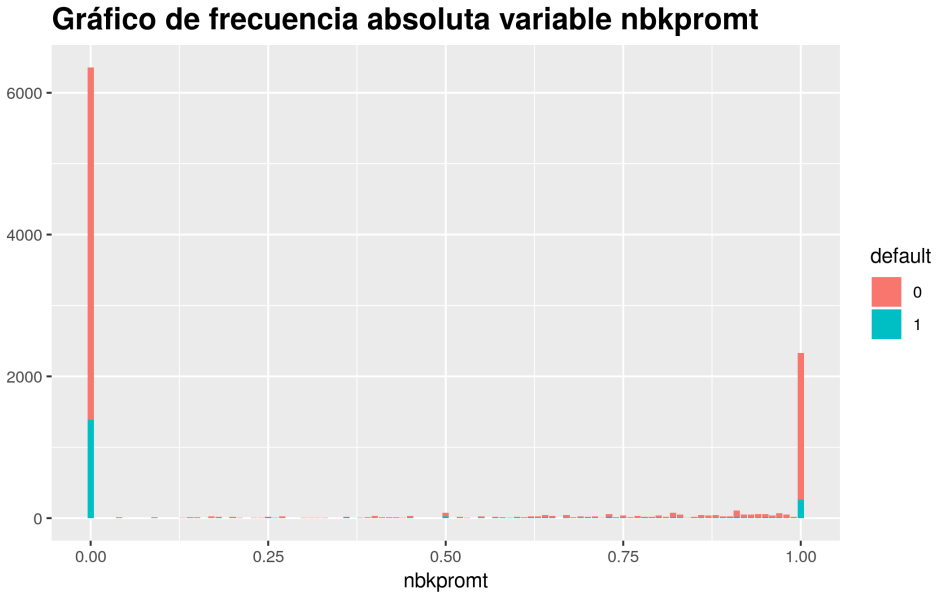
tasa de default es del 57 %, y por último cuando toma el valor de 3 la tasa de default es del 76 %. Dicho lo anterior, esta variable pareciera que tiene poder predictivo como se muestra en la siguiente gráfica.



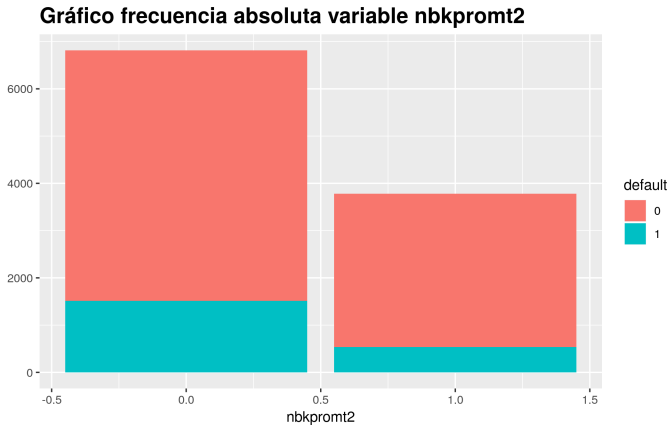

En el siguiente gráfico se muestra la frecuencia absoluta de la variable bkpromt, no es muy claro si conforme más cercana a 1 es esta variable, en términos relativos disminuye o aumenta la tasa de default.



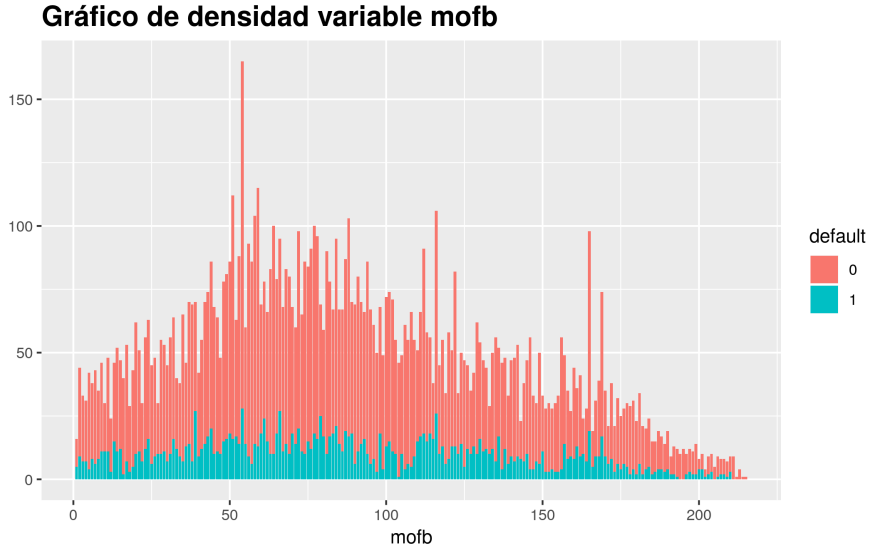
En el siguiente gráfico se muestra la frecuencia absoluta de la variable nbkpromt, notemos que la mayoría de cuentas se concentran en el valor uno o en el valor cero, por ello se propone construir la variable nbkpromt2, la cual toma el valor de cero cuando nbkpromt es menor o igual a 0.5, y 1 en caso contrario.



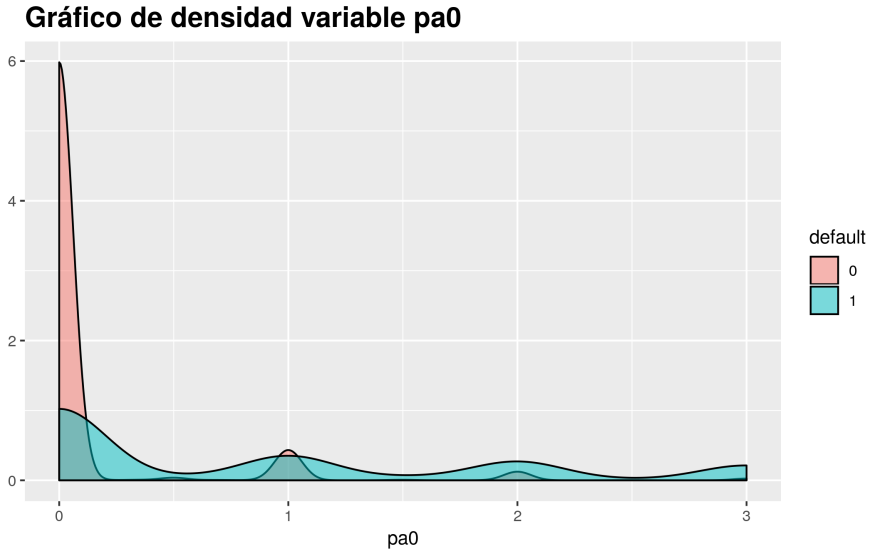

En el siguiente gráfico se muestra la variable nbkpromt2, descrita anteriormente, notemos que cuando esta variable toma el valor de 0 la tasa de default es del 23 %, mientras que cuando toma el valor de 1 la tasa de default es del 14 %. Dicho lo anterior, tenemos evidencia a favor de que esta variable puede tener poder predictivo, para así ser incorporada en el modelo.



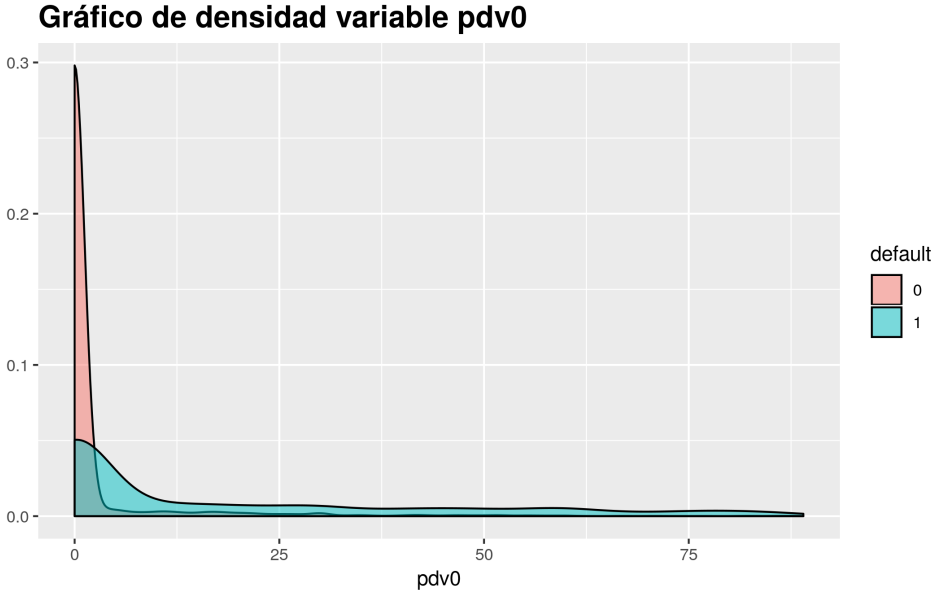
En el siguiente gráfico se muestra la frecuencia absoluta de la variable mofb, notemos que no hay una interpretación clara de si entré mayor es esta variable, en términos relativos, disminuye la tasa de default.



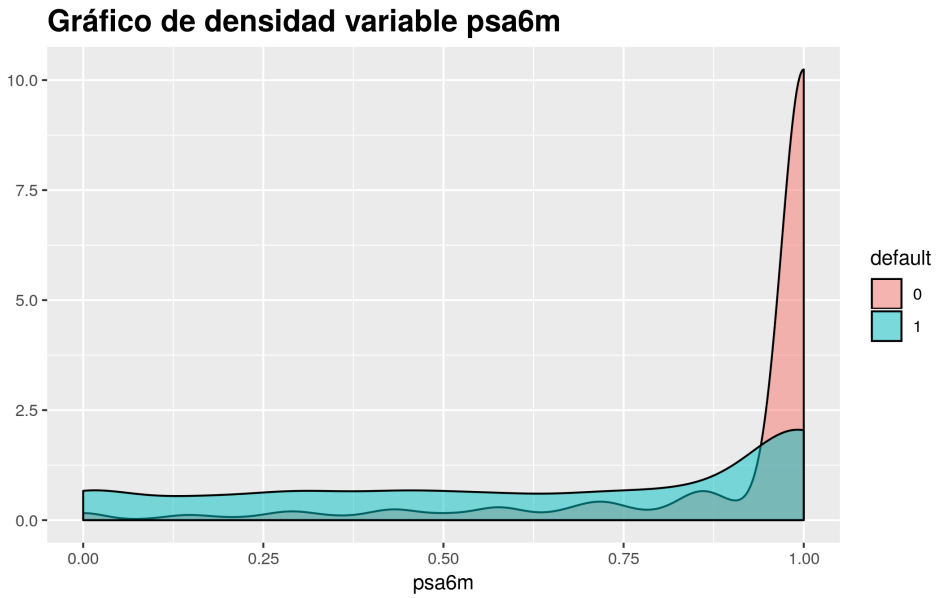

En el siguiente gráfico de densidad se muestra la distribución de defaults a través del rango de la variable pa0, notemos que la mayoría de las cuentas que no fueron default se concentran alrededor del valor cero, mientras que las cuentas que tuvieron default se distribuyen a lo largo de esta variable.



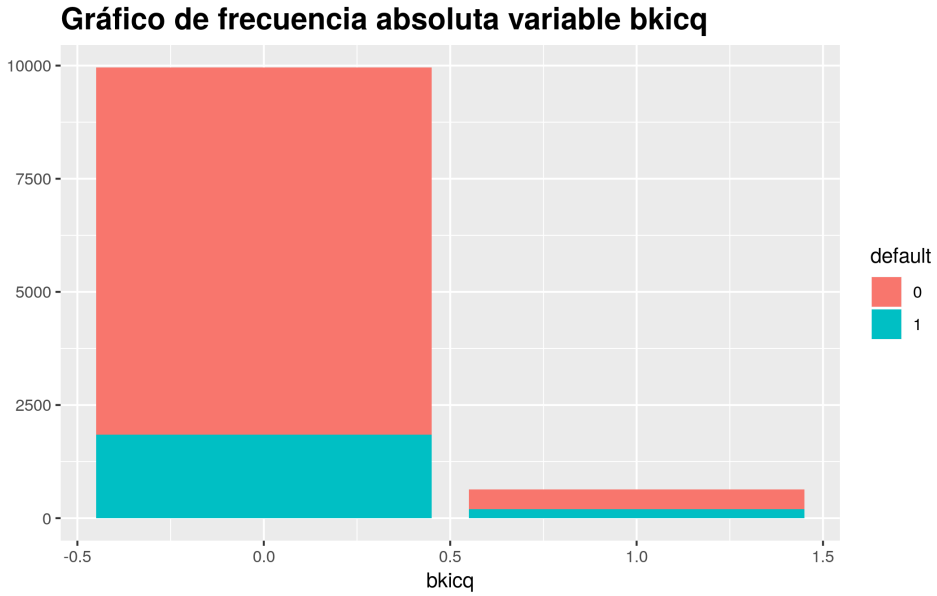
Al igual que en el gráfico anterior, en el siguiente gráfico de densidad se muestra la distribución de defaults a través del rango de la variable pdv0, notemos que la mayoría de las cuentas que no fueron default se concentran al rededor del valor cero, mientras que las cuentas que tuvieron default se distribuyen a lo largo de esta variable.



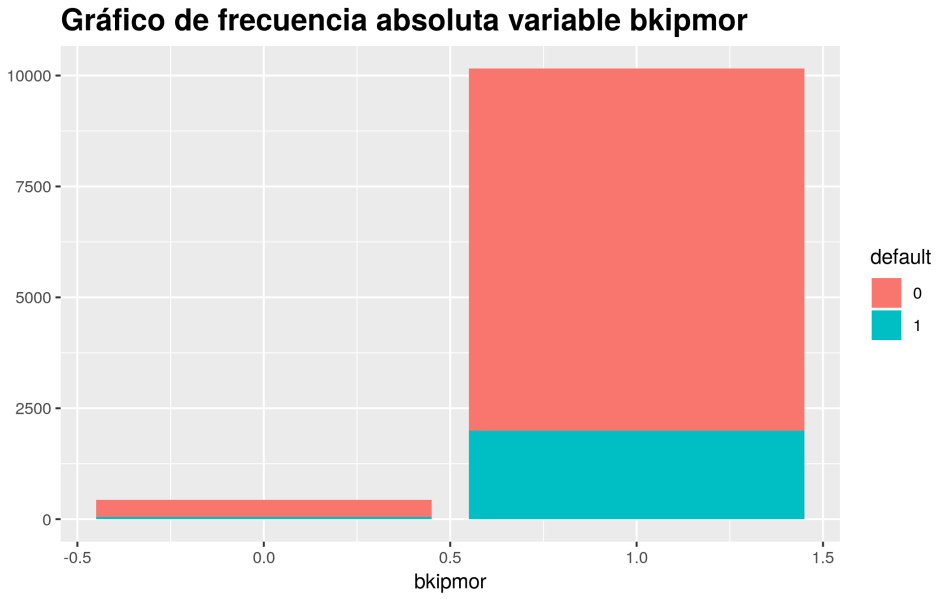

De igual manera que en el gráfico anterior, en el siguiente gráfico de densidad se muestra la distribución de defaults a través del rango de la variable psa6m. Se puede observar que la mayoría de las cuentas que no fueron default se concentran alrededor del valor uno, mientras que las cuentas que tuvieron default se distribuyen de manera muy similar en el rango de esta variable.




En el siguiente gráfico se muestra la frecuencia absoluta de la variable bkicq, notemos que aunque la tasa de default cuando esta variable toma el valor de cero es del 19 %, y de 38 % cuando toma el valor de 1. El número de cuentas donde esta variable toma el valor de 1 es poco menos del 5 % de la población total, lo que podría ocasionar que esta variable por sí sola no tenga poder predictivo.



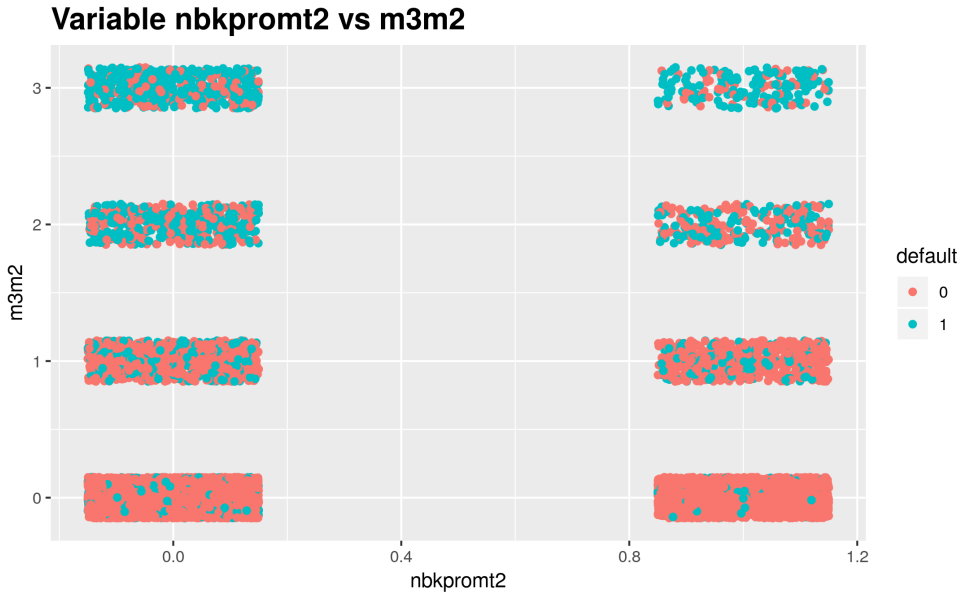
En el siguiente gráfico se muestra la frecuencia absoluta de la variable bkipmor. Análogamente al caso anterior, notemos que el número de cuentas cuando esta variable toma el valor de cero representa poco menos del 4 % de la población total, lo que podría ocasionar que esta variable por sí sola no tenga poder predictivo.



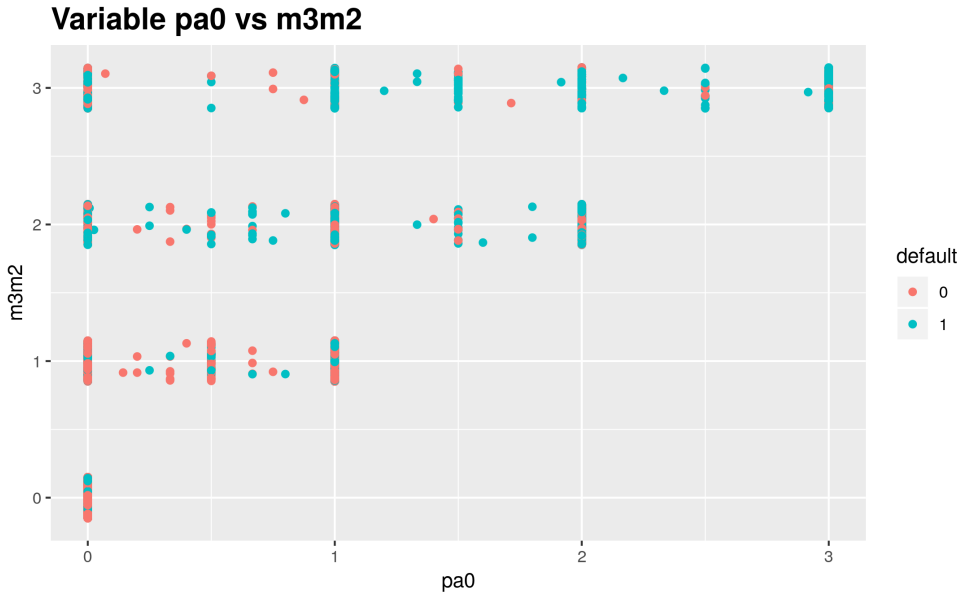

## 2.5. Efectos cruzados

A continuación se habla brevemente de los principales efectos cruzados encontrados durante el análisis exploratorio de datos.

En el siguiente gráfico se muestra la dispersión entre las variables nbkpromt2 y m3m2, notemos que, como ya se había mencionado anteriormente, entre mayor es el valor que toma la variable m3m2, mayor es la tasa de default, pero además la tasa de default está influenciada por el valor que toma la variable nbkpromt2, pues cuando m3m2 es mayor o igual a 3, y nbkpromt2 vale cero esta es del 70 % aproximadamente, mientras que cuando m3m2 vale 3 y nbkpromt2 vale uno, la tasa de default es del 82 % aproximadamente, por ejemplo.

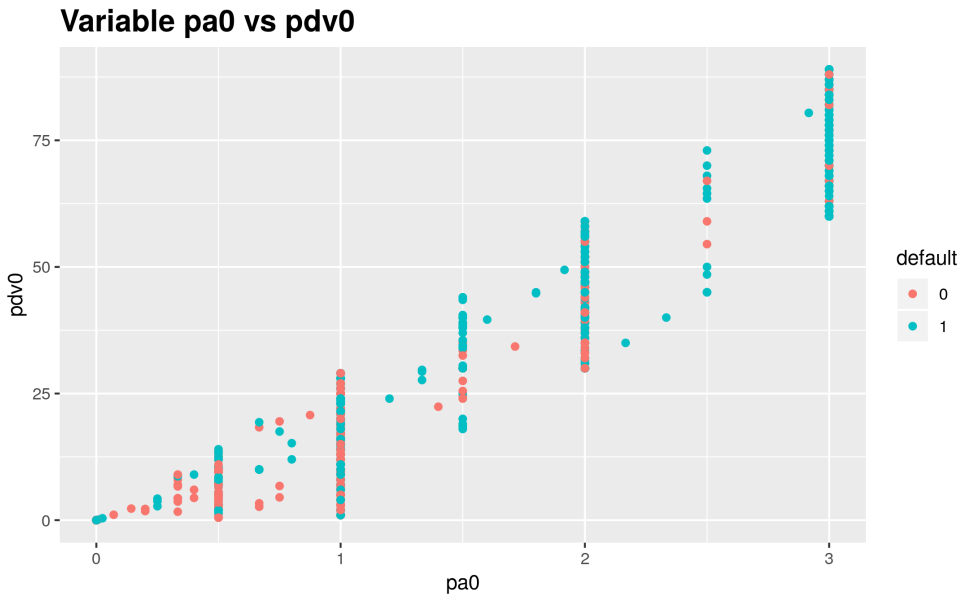



En el siguiente gráfico se muestra la dispersión entre las variables pa0 y m3m2, pareciera que entre mayor son ambas variables se observa una mayor cantidad de cuentas en default.



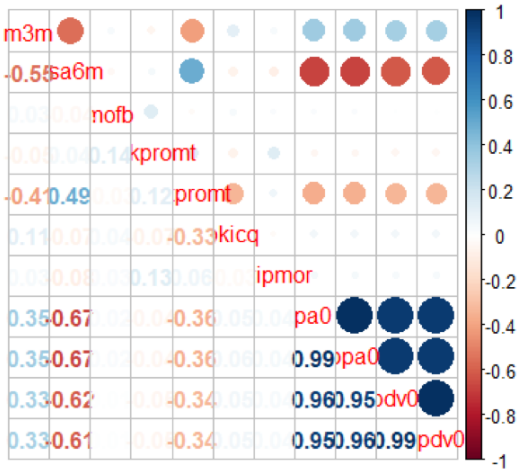

En el siguiente gráfico se muestra la dispersión entre las variables pa0 y pdv0, al igual que en el caso anterior, pareciera que entre mayor son ambas variables

se observa una mayor cantidad de cuentas en default.



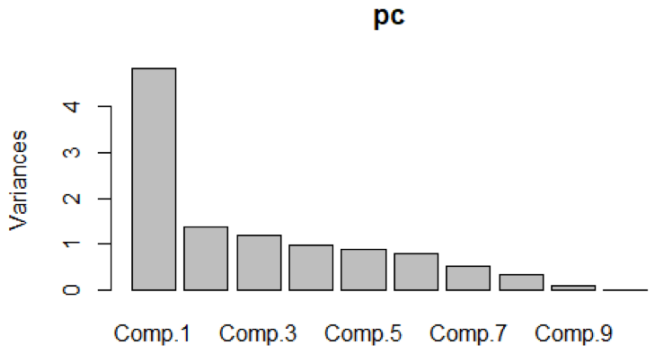

## 2.6. PCA: Análisis de componentes principales

Una vez realizado el análisis exploratorio de datos (EDA), se muestra las correlaciones entre las covariables seleccionadas para el análisis de componentes principales (ACP):

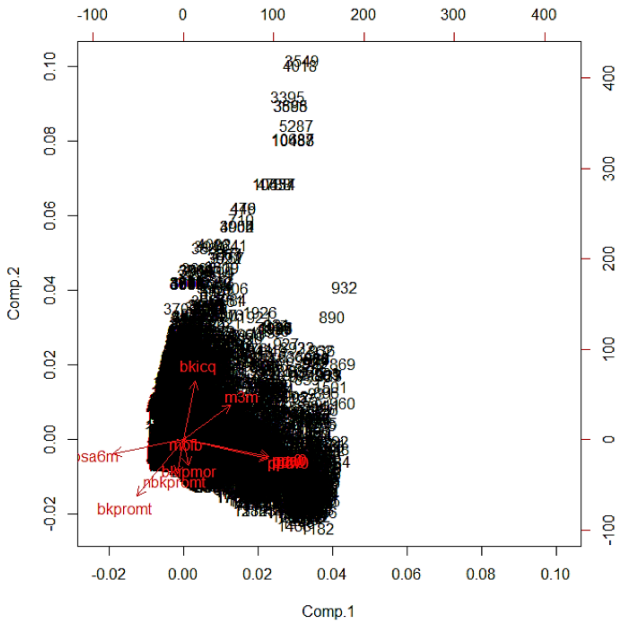


Como se puede apreciar, varias de estas variables están altamente correlacionadas entre sí, por lo que el ACP es una herramienta idónea para reducir la dimensión del problema y obtener medidas resumen. Se procedió a realizar un análisis de componentes principales (PCA) de la matriz de varianzas y covarianzas.

Podemos ver que basta con utilizar las primeras tres componentes principales para obtener casi el 70 % de la variación total en los datos. El resto del análisis se basará únicamente en las componentes PC1, PC2, PC3.




Proyecciones de los Componente Principales




La gráfica anterior muestra la proyección de los datos sobre los planos de las combinaciones de las dos primeras componentes. Se puede notar que la primer

componente pondera prácticamente con la misma magnitud a las variable pa0. Así mismo las variables psa6m y bkpromt tienen peso negativo. De esta manera, la primer componente principal sería casi equivalente al promedio de las anteriores covariables. La segunda componente está determinada principalmente por m3m y bkcq. La tercer componente está determinada principalmente por las variables mofb, nbkpromt y bkipmor.

Con base en el PCA se pudo distinguir a cuatro grupos principales de variables:

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

Tomando como base a dichos grupos, se eligió a las variables m3m y bkpromt.

Las variables que se seleccionaron para ser parte del modelo después del análisis fueron las siguientes:

| **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: $(nbkpromt2 + .01) * ifelse(\text{m3m} \geq 3, 3, \text{m3m})$ |

# 3. Modelado e Implementación

Una vez seleccionadas las variables, se propone probar un modelo Bernoulli con liga logística, liga probit, liga log-log y liga clog-log. El modelo en consideración tendrá la siguiente estructura seleccionando un modelo de regresión Bernoulli:

$$y_i \mid \mu_i \sim Ber(\mu_i = \theta), \theta \in (0, 1) \tag{1}$$

$$y_i \mid \mu_i \sim Bernoulli(\theta) \tag{2}$$

Con función de densidad condicional:

$f(y_i|\mu_i) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} I_{\{0,1\}}(y_i)$,

donde:

1. $\eta = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \beta_4 X_3 + \beta_5 X_4 + \beta_6 X_5$.
2. $\Phi_i = 1$.
3. $b(y_i, \Phi_i) = b(y_i)$.
4. $\theta_i = log(\mu_i/(1-\mu_i))$.
5. $a(\theta_i) = log(1 + e^{\theta_i})$.

La función de densidad a priori para cada coeficiente se consideró normal no informativa de la siguiente manera:

$$\beta_i \sim N(0, 0.01), \tag{3}$$

donde $\tau = 1/\sigma^2 = 0.01$.

A continuación, se mostrarán las fórmulas correspondientes a las siguientes ligas:

Logística:

$$\theta = \frac{1}{1+e^{\eta}}. \tag{4}$$

Probit:

$$\theta = \Phi(\eta). \tag{5}$$

log-log:

$$\theta = log(-log(\eta)). \tag{6}$$

log-log complementaria:

$$\theta = log(-log(1-\eta)). \tag{7}$$

## Consideraciones de los modelos propuestos

Para la realización del modelo se prosiguió de la siguiente manera:

1. Dada las naturaleza de las variables, todas fueron tratadas como numéricas. Y además, fueron estandarizadas, es decir, a cada observación se le restó su respectiva media y se dividió entre su desviación estándar.

2. Las variables bkicq, bkipmor fueron descartadas desde un principio conforme se describe en el análisis exploratorio de datos.

3. Para las demás variables se realizaron corridas de los 4 modelos propuestos, y se fueron descartando según si el intervalo de probabilidad contenía el valor 0 o no. Pues este nos indica con cierto nivel de probabilidad el valor estimado de cada valor $\beta$ puede ser de cero, es decir, es no significativa. Bajo este criterio fueron descartadas las variables mofb, pdv0, psa6m, y las interacciones papdv0 y pam3.

4. Las variables que resultaron significativas en los 4 modelos fueron las dos variables transformadas m3m2 y nbkpromt2, las variables bkpromt y pa0, y la interacción nbm3 (las variables de interacción también fueron estandarizadas).

5. Para implementar cada modelo se empleo el software estadístico **R** en conjunto con **JAGS** *(Just Another Gibbs Sampler)*. Se realizó cada simulación con **20,000** iteraciones, **2** cadenas y un adelgazamiento de **5** para los cuatro modelos correspondientes a cada una de las funciones liga.

   Se alcanzó convergencia con dicho número de iteraciones al observar las cadenas de Markov para cada uno de los modelos, para lo cual incluimos una gráfica en la sección de "Anexos" al final del presente proyecto. Así mismo, soportando esta misma idea, los coeficientes y el DIC empezaron a estabilizar desde las 10,000 iteraciones.

### Modelo Bernoulli con liga logística

A continuación se presentan el estimador bayesiano de las $\beta$'s del modelo propuesto bajo la función de pérdida cuadrática, así mismo se muestra el intervalo de probabilidad del 2.5 % al 97.5 %.

| Nombre | Coeficiente | mean | 2.5 % | 97.5 % | Rhat |
|---|---|---|---|---|---|
| Intercepto | $\beta_1$ | -1.72 | -1.78 | -1.66 | 1.00 |
| m3m2 | $\beta_2$ | 0.52 | 0.44 | 0.59 | 1.00 |
| nbkpromt2 | $\beta_3$ | -0.28 | -0.36 | -0.20 | 1.00 |
| bkpromt | $\beta_4$ | -0.38 | -0.43 | -0.32 | 1.00 |
| pa0 | $\beta_5$ | 0.41 | 0.35 | 0.47 | 1.00 |
| nbm3 | $\beta_6$ | 0.09 | 0.03 | 0.16 | 1.00 |
| DIC | - | 8027.20 | 8021.93 | 8035.23 | 1.00 |

**Modelo Bernoulli con liga probit**

A continuación, se presentan el estimador bayesiano de las $\beta$'s del modelo propuesto bajo la función de pérdida cuadrática, así mismo se muestra el intervalo de probabilidad del 2.5 % al 97.5 %.

| Nombre | Coeficiente | mean | 2.5 % | 97.5 % | Rhat |
|---|---|---|---|---|---|
| Intercepto | $\beta_1$ | -1.01 | -1.04 | -0.98 | 1 |
| m3m2 | $\beta_2$ | 0.30 | 0.26 | 0.35 | 1 |
| nbkpromt2 | $\beta_3$ | -0.15 | -0.19 | -0.11 | 1 |
| bkpromt | $\beta_4$ | -0.22 | -0.25 | -0.19 | 1 |
| pa0 | $\beta_5$ | 0.24 | 0.20 | 0.27 | 1 |
| nbm3 | $\beta_6$ | 0.05 | 0.01 | 0.08 | 1 |
| DIC - | | 8008.86 | 8003.68 | 8016.93 | 1 |

**Modelo Bernoulli con liga log-log**

A continuación, se presentan el estimador bayesiano de las $\beta$'s del modelo propuesto bajo la función de pérdida cuadrática, así mismo se muestra el intervalo de probabilidad del 2.5 % al 97.5 %.

| Nombre | Coeficiente | mean | 2.5 % | 97.5 % | Rhat |
|---|---|---|---|---|---|
| Intercepto | $\beta_1$ | 0.56 | 0.53 | 0.59 | 1 |
| m3m2 | $\beta_2$ | -0.29 | -0.34 | -0.24 | 1 |
| nbkpromt2 | $\beta_3$ | 0.11 | 0.08 | 0.14 | 1 |
| bkpromt | $\beta_4$ | 0.25 | 0.21 | 0.28 | 1 |
| pa0 | $\beta_5$ | -0.28 | -0.32 | -0.23 | 1 |
| nbm3 | $\beta_6$ | -0.01 | -0.05 | 0.03 | 1 |
| DIC | - | 7952.35 | 7947.62 | 7960.58 | 1 |

**Modelo Bernoulli con liga log-log complementaria**

A continuación, se presentan el estimador bayesiano de las $\beta$'s del modelo propuesto bajo la función de pérdida cuadrática, así mismo se muestra el intervalo de probabilidad del 2.5 % al 97.5 %.

| Nombre | Coeficiente | mean | 2.5 % | 97.5 % | Rhat |
|---|---|---|---|---|---|
| Intercepto | $\beta_1$ | -1.83 | -1.89 | -1.78 | 1 |
| m3m2 | $\beta_2$ | 0.44 | 0.39 | 0.50 | 1 |
| nbkpromt2 | $\beta_3$ | -0.27 | -0.34 | -0.20 | 1 |
| bkpromt | $\beta_4$ | -0.21 | -0.25 | -0.18 | 1 |
| pa0 | $\beta_5$ | 0.24 | 0.20 | 0.28 | 1 |
| nbm3 | $\beta_6$ | 0.11 | 0.06 | 0.15 | 1 |
| DIC | - | 8128.63 | 8123.85 | 8137.26 | 1 |

# 4. 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:

$$log(-log(\mu_i)) = \eta_i = x_i\beta \iff \mu_i = exp\{-exp[x_i\beta]\}, \tag{8}$$

Sea,

$$x_j = x_i + 1. \tag{9}$$

A la variable explicativa le agregamos una unidad tal que $x_i = 1, x_{i1}, x_{i2}$ con dos variables explicativas.

$$x_j = 1, x_{i1} + 1, x_{i2}, \tag{10}$$

Así,

$$log(-log(\mu_j)) = x_j\beta = \beta_0 + \beta_1(x_{i1} + 1) + \beta_2 x_{i2}. \tag{11}$$

La anterior tiene un $\beta_1$ más que al evaluar en el individuo i:

$$log(-log(\mu_i)) = \eta_i = x_i\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \tag{12}$$

Así, la diferencia de logaritmos da:

$$log(-log(\mu_j)) - log(-log(\mu_i)) = \beta_1, \tag{13}$$

Si y sólo si,

$$log\left(\frac{-log(\mu_j)}{-log(\mu_i)}\right) = log\left(\frac{log(\mu_j)}{log(\mu_i)}\right), \tag{14}$$

Si y sólo si,

$$\frac{log(\mu_j)}{log(\mu_i)} = e^{\beta_1}. \tag{15}$$

A $e^{\beta_1}$ se le conoce como cociente de logaritmos. Como el logaritmo transforma de [0,1] a negativos, los signos de esta liga son distintos a los de otras. Es decir, tienen el signo contrario. Por lo tanto, indica cómo cambia el cociente de los logaritmos.

## 4.1. Algunas Predicciones

Correctamente clasificados (umbral 62.5 % para clasificarse como default):

| | y | y_hat | m3m2 | nbkpromt2 | bkpromt | pa0 | nbm3 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1.00 | 0.00 | 1.00 | 0.94 | 0.00 | 0.00 |
| 2 | 1 | 1.00 | 1.00 | 0.00 | 0.93 | 0.00 | 0.01 |
| 3 | 0 | 0.00 | 2.00 | 0.00 | 0.50 | 2.00 | 0.02 |
| 4 | 0 | 0.00 | 3.00 | 0.00 | 0.52 | 2.00 | 0.03 |

Incorrectamente clasificados (umbral 62.5 % para clasificarse como default):

| | y | y_hat | m3m2 | nbkpromt2 | bkpromt | pa0 | nbm3 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0.00 | 1.00 | 0.00 | 0.21 | 0.00 | 0.01 |
| 2 | 1 | 0.00 | 2.00 | 0.00 | 0.77 | 1.00 | 0.02 |
| 3 | 1 | 0.00 | 3.00 | 0.00 | 0.54 | 0.00 | 0.03 |

Correctamente clasificados (umbral 12.5 % para clasificarse como default):

| | y | y_hat | m3m2 | nbkpromt2 | bkpromt | pa0 | nbm3 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 1.00 | 0.00 | 1.00 | 0.94 | 0.00 | 0.00 |
| 2 | 0 | 0.00 | 2.00 | 1.00 | 0.98 | 2.00 | 2.02 |
| 3 | 0 | 0.00 | 2.00 | 0.00 | 0.67 | 2.00 | 0.02 |
| 4 | 0 | 0.00 | 3.00 | 0.00 | 0.90 | 1.50 | 0.03 |

Incorrectamente clasificados (umbral 12.5 % para clasificarse como default):

| | y | y_hat | m3m2 | nbkpromt2 | bkpromt | pa0 | nbm3 |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 1.00 | 0.00 | 0.00 | 0.78 | 0.00 | 0.00 |
| 2 | 0 | 1.00 | 1.00 | 1.00 | 1.00 | 0.50 | 1.01 |
| 3 | 0 | 1.00 | 2.00 | 1.00 | 0.87 | 1.00 | 2.02 |
| 4 | 0 | 1.00 | 3.00 | 0.00 | 0.83 | 0.00 | 0.03 |

# 5. Conclusiones

Se cumplió el objetivo de obtener un «modelo de score» basado en la probabilidad de default con resultados razonables mediante un modelo lineal generalizado con enfoque bayesiano. 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 si cae en default.

# Referencias


[1] Nicky Best Dave Lunn David Spiegelhalter, Andrew Thomas. *OpenBUGS User Manual*. Cambdridge University, March 2014.

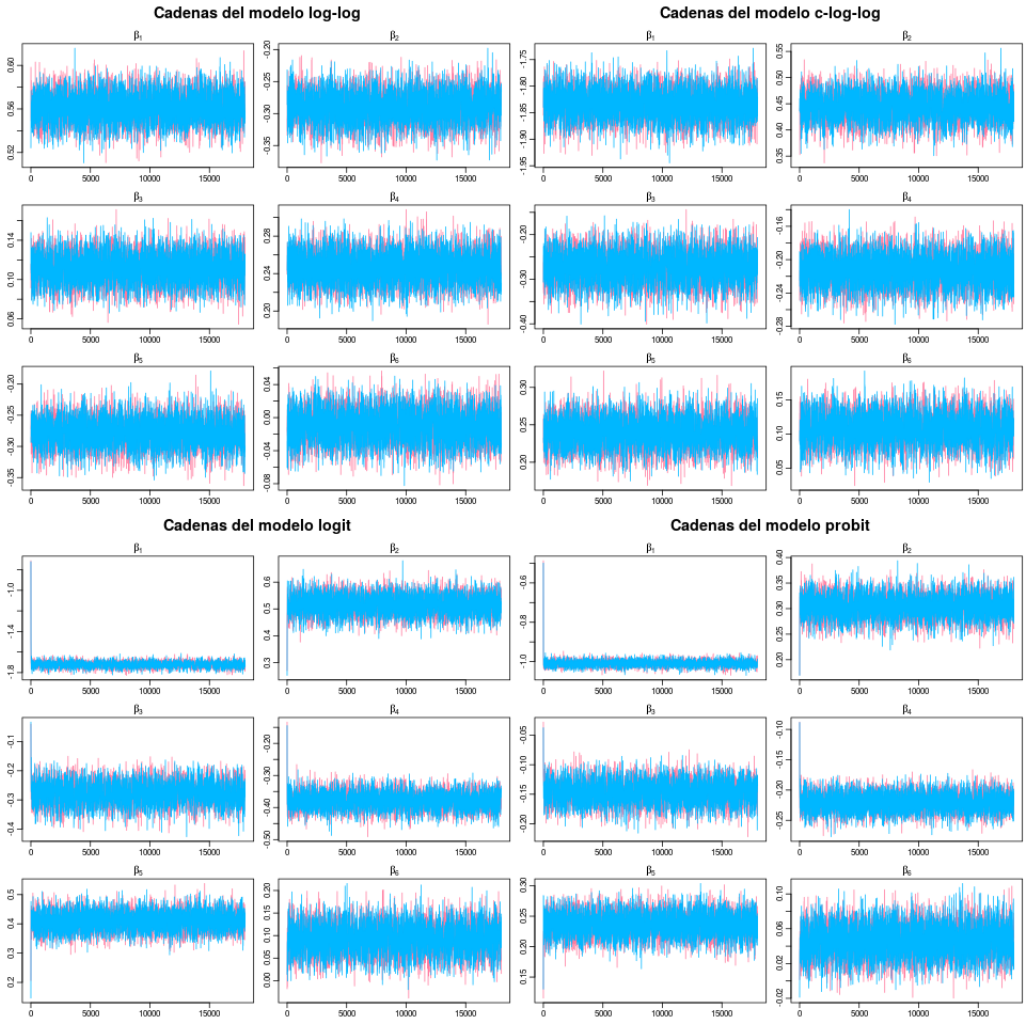
[2] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[3] 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.

[4] Luis E. Nieto-Barajas. Notas del curso de regresión avanzada. 2019.

# 6. Apéndice

## 6.1. Convergencia de las Cadenas



## 6.2. Código R

```
#Carga de Librerías
library(tidyverse)
library(R2jags)
library(R2OpenBUGS)
library(plotly)
library(rstan)
library(mcmcplots)
knitr::opts_chunk$set(echo = FALSE,
warning = FALSE,
messages = FALSE,
include = TRUE,
results = "show")
#funciones útiles
print.latex <- function(input){
if(is.null(knitr::opts_knit$get("rmarkdown.pandoc.to"))){
input %>% data.frame
} else {
if(knitr::opts_knit$get("rmarkdown.pandoc.to") == 'latex'){
input %>% xtable::xtable(.) %>% print(comment=F, type="latex")
} else {
input %>% data.frame
}
}
}
prepare.latex <- function(latex=T){
if(latex){
knitr::opts_chunk$set(results="asis")
} else{
knitr::opts_chunk$set(results="show")
}
}
prob<-function(x){
out<-min(length(x[x>0])/length(x),length(x[x<0])/length(x))
out
}
muestra <- read.csv("muestra.csv", header=T)
muestra <- muestra[-which(muestra$pdv0<0),]
#Construcción de variables:
muestra$m3m2<-ifelse(muestra$m3m>=3,3,muestra$m3m)
muestra$nbkpromt2<-ifelse(muestra$nbkpromt<=.5,0,1)
muestra$nbm3<-(muestra$nbkpromt2+.01)*muestra$m3m2
muestra$pam3<-muestra$pa0*muestra$m3m2
muestra$papd<-muestra$pa0*muestra$pdv0
## división de entrenamiento y validación
```

```
train_size <- (muestra %>% nrow * 0.7) %>% floor
test_size  <- muestra %>% nrow - train_size
set.seed(175904)
train      <- muestra[sample(1:nrow(muestra),train_size,replace =FALSE),]
test       <- muestra[sample(1:nrow(muestra),test_size,replace =FALSE),]
## preparaciones para los cálculos de modelos
N <- nrow(train)
M <- nrow(test)
inits <- function() { list(beta = rep(0, 6)) }
x_vars <- c("m3m2","nbkpromt2","bkpromt","pa0","nbm3")
x<-train[,x_vars]
# x[train_size:(train_size+test_size),] <- NA
data    <- list("n" = N, "m" = M, "y" = train$y,
"x"=x%>%scale, "xf"=test[x_vars]%>%scale)
pars    <- c("beta", "yf2")
logit_model <- jags.parallel(data, inits, pars, model.file = "final-dbin-logit.txt",
n.iter = 20000, n.chains = 2, n.burnin = 2000, n.thin = 5)
cloglog_model <- jags.parallel(data, inits, pars, model.file = "final-dbin-clog.txt",
n.iter = 20000, n.chains = 2, n.burnin = 2000, n.thin = 5)
loglog_model <- jags.parallel(data, inits, pars, model.file = "final-dbin-loglog.txt",
n.iter = 20000, n.chains = 2, n.burnin = 2000, n.thin = 5)
probit_model <- jags.parallel(data, inits, pars, model.file = "final-dbin-probit.txt",
n.iter = 20000, n.chains = 2, n.burnin = 2000, n.thin = 5)
##eda
muestra %>% summary
muestra %>% scale %>% boxplot
##geda
#Agrupamiento de la variable m3m.
p1<-ggplot(train, aes(m3m,fill=as.factor(y)))+geom_bar(stat = "count")+labs(fill = "default'
p2<-ggplot(train, aes(m3m2,fill=as.factor(y)))+geom_bar(stat = "count")+labs(fill = "default
p3<-ggplot(train,aes(x=psa6m,fill=as.factor(y)))+geom_density(alpha = 0.5)+labs(fill = "defa
#Agrupamiento de la variable m3m.
p4<-ggplot(train, aes(mofb,fill=as.factor(y)))+geom_bar(stat = "count")+labs(fill = "default
################################################################################
#Agrupamiento de la variable nbkpromt2
p5<-ggplot(train,aes(nbkpromt,fill=as.factor(y)))+geom_bar(stat = "count")+labs(fill = "defa
p6<-ggplot(train, aes(nbkpromt2,fill=as.factor(y)))+geom_bar(stat = "count")+labs(fill = "de
#################
#################
p7<-ggplot(train, aes(x=nbkpromt2+runif(1:nrow(train),-0.15,0.15),y=m3m2+runif(1:nrow(train)
################################################################################
################################################################################
p8<-ggplot(train, aes(pa0,fill=as.factor(y)))+geom_density(alpha = 0.5)+labs(fill = "default
p9<-ggplot(train, aes(ppa0,fill=as.factor(y)))+geom_density(alpha = 0.5)+labs(fill = "defaul
#ggplot(train, aes(ppa0,fill=as.factor(y)))+geom_density()+labs(fill = "default")+xlab("nbkp
p10<-ggplot(train, aes(x=pa0,y=m3m2+runif(1:nrow(train),-0.15,0.15),color=as.factor(y)))+geo
```

```
#Podría funcionar
p10<-ggplot(train, aes(pdv0,fill=as.factor(y)))+geom_density(alpha = 0.5)+labs(fill = "defau
p11<-ggplot(train, aes(ppdv0,fill=as.factor(y)))+geom_density(alpha = 0.5)+labs(fill = "defa
p12<-ggplot(train, aes(x=pa0,y=pdv0,color=as.factor(y)))+geom_point()+labs(color = "default'
##############################################################################################
##############################################################################################
p13<-ggplot(train,aes(bkpromt,fill=as.factor(y)))+geom_bar(stat = "count")+labs(fill = "defa
p14<-ggplot(train, aes(bkicq,fill=as.factor(y)))+geom_bar(stat = "count")+labs(fill = "defau
p15<-ggplot(train, aes(bkipmor,fill=as.factor(y)))+geom_bar(stat = "count")+labs(fill = "def
p16<-ggplot(train, aes(x=pa0,y=m3m2+runif(1:nrow(train),-0.15,0.15),color=as.factor(y)))+geo
ggplot(train, aes(nbm3,fill=as.factor(y)))+geom_bar(stat = "count")+labs(fill = "default")+x
p1
p2
p3
p4
p5
p6
p7
p8
p9
p10
p11
p12
p13
p14
p15
p16
ggsave(filename = "img/m3m.png", plot = p1)
ggsave(filename = "img/m3m2.png", plot = p2)
ggsave(filename = "img/psa6m.png", plot = p3)
ggsave(filename = "img/mofb.png", plot = p4)
ggsave(filename = "img/nbkpromt.png", plot = p5)
ggsave(filename = "img/nbkpromt2.png", plot = p6)
ggsave(filename = "img/nbkpromt2vsm3m2.png", plot = p7)
ggsave(filename = "img/pa0.png", plot = p8)
ggsave(filename = "img/ppa0.png", plot = p9)
ggsave(filename = "img/pdv0.png", plot = p10)
ggsave(filename = "img/ppdv0.png", plot = p11)
ggsave(filename = "img/pa0vspdv0.png", plot = p12)
ggsave(filename = "img/bkpromt.png", plot = p13)
ggsave(filename = "img/bkicq.png", plot = p14)
ggsave(filename = "img/bkipmor.png", plot = p15)
ggsave(filename = "img/pa0vsm3m2.png", plot = p16)
## train, test sets analysis
paste("Observaciones utilizadas para el conjunto de entrenamiento (~70%):", train_size, "\n'
paste("Observaciones utilizadas para el conjunto de para prueba (~30%):", test_size, "\n") %
```

```
paste('Media de la "y" en el conjunto original:', mean(muestra$y) %>% round(digits=4), "\n")
paste('Media de la "y" en el conjunto de entrenamiento:', mean(train$y) %>% round(digits=4),
paste('Media de la "y" en el conjunto de pruebas', mean(test$y) %>% round(digits=4), "\n") %
## modelado, analisis
print(logit_model$BUGSoutput$summary[1:7,])
cat(paste('\nDIC = ',
logit_model$BUGSoutput$DIC))
print(cloglog_model$BUGSoutput$summary[1:7,])
cat(paste('\nDIC = ',
cloglog_model$BUGSoutput$DIC))
cloglog_model$BUGSoutput$summary[1:7,1] %>% sum
print(loglog_model$BUGSoutput$summary[1:7,])
cat(paste('\nDIC = ',
cloglog_model$BUGSoutput$DIC))
loglog_model$BUGSoutput$summary[1:7,1] %>% sum
print(probit_model$BUGSoutput$summary[1:7,])
cat(paste('\nDIC = ',
probit_model$BUGSoutput$DIC))
probit_model$BUGSoutput$summary[1:7,1] %>% sum
## análisis de convergencia
# png("img/traceplot-logit.png")
traplot(logit_model, parms = c("beta"), auto.layout = TRUE, greek=T,
style=c("plain"), plot.title="Cadenas del modelo logit")
# png("img/traceplot-logit.png")
traplot(cloglog_model, parms = c("beta"), auto.layout = TRUE, greek=T,
style=c("plain"), plot.title="Cadenas del modelo c-log-log")
# png("img/traceplot-logit.png")
traplot(loglog_model, parms = c("beta"), auto.layout = TRUE, greek=T,
style=c("plain"), plot.title="Cadenas del modelo log-log")
# png("img/traceplot-logit.png")
traplot(probit_model, parms = c("beta"), auto.layout = TRUE, greek=T,
style=c("plain"), plot.title="Cadenas del modelo probit")
# dev.off()
## comparación de los modelos
loglog_txt <- c(paste0(loglog_model$BUGSoutput$summary[1,1]%>%round(2),
" (",
loglog_model$BUGSoutput$summary[1,1]%>%quantile(c(0.0025),names=F)%>%round(2),
", ",
loglog_model$BUGSoutput$summary[1,1]%>%quantile(c(0.9725),names=F)%>%round(2),
")"),
paste0(loglog_model$BUGSoutput$summary[2,1]%>%round(2),
" (",
loglog_model$BUGSoutput$summary[2,1]%>%quantile(c(0.0025),names=F)%>%round(2),
", ",
loglog_model$BUGSoutput$summary[2,1]%>%quantile(c(0.9725),names=F)%>%round(2),
")"),
```

```
paste0(loglog_model$BUGSoutput$summary[3,1]%>%round(2),
" (",
loglog_model$BUGSoutput$summary[3,1]%>%quantile(c(0.0025),names=F)%>%round(2),
", ",
loglog_model$BUGSoutput$summary[3,1]%>%quantile(c(0.9725),names=F)%>%round(2),
")"),
paste0(loglog_model$BUGSoutput$summary[4,1]%>%round(2),
" (",
loglog_model$BUGSoutput$summary[4,1]%>%quantile(c(0.0025),names=F)%>%round(2),
", ",
loglog_model$BUGSoutput$summary[4,1]%>%quantile(c(0.9725),names=F)%>%round(2),
")"),
paste0(loglog_model$BUGSoutput$summary[5,1]%>%round(2),
" (",
loglog_model$BUGSoutput$summary[5,1]%>%quantile(c(0.0025),names=F)%>%round(2),
", ",
loglog_model$BUGSoutput$summary[5,1]%>%quantile(c(0.9725),names=F)%>%round(2),
")"),
paste0(loglog_model$BUGSoutput$summary[6,1]%>%round(2),
" (",
loglog_model$BUGSoutput$summary[6,1]%>%quantile(c(0.0025),names=F)%>%round(2),
", ",
loglog_model$BUGSoutput$summary[6,1]%>%quantile(c(0.9725),names=F)%>%round(2),
")"))
```

### 6.3. Código BUGS Logit

```
model
{
#Likelihood
for (i in 1:n) {
y[i] ~ dbin(p[i],1)
logit(p[i])<-beta[1] + beta[2]*x[i,1] + beta[3]*x[i,2] + beta[4]*x[i,3] + beta[5]*x[i,4] + b
}
#Priors
for (j in 1:6) {
beta[j] ~ dnorm(0,0.001)
}


#Prediction 1
for (i in 1:n) {
yf1[i] ~ dbin(p[i],1)
}
#Prediction 2
for (i in 1:m) {
```

```
logit(pf[i]) <- beta[1] + beta[2]*xf[i,1] + beta[3]*xf[i,2] + beta[4]*xf[i,3] + beta[5]*xf[i
yf2[i] ~ dbin(pf[i],1)
}
}
```

## 6.4. Código BUGS Probit

```
model
{
#Likelihood
for (i in 1:n) {
y[i] ~ dbin(p[i],1)
eta[i]<-beta[1] + beta[2]*x[i,1] + beta[3]*x[i,2] + beta[4]*x[i,3] + beta[5]*x[i,4] + beta[6
p[i] <- phi(eta[i])
}
#Priors
for (j in 1:6) {
beta[j] ~ dnorm(0,0.001)
}

#Prediction 1
for (i in 1:n) {
yf1[i] ~ dbin(p[i],1)
}
#Prediction 2
for (i in 1:m) {
etaf[i] <- beta[1] + beta[2]*xf[i,1] + beta[3]*xf[i,2] + beta[4]*xf[i,3] + beta[5]*xf[i,4] +
pf[i] <- phi(etaf[i])
yf2[i] ~ dbin(pf[i],1)
}
}
```

## 6.5. Código BUGS Log Log

```
model
{
#Likelihood
for (i in 1:n) {
y[i] ~ dbin(p[i],1)
#Liga log-log
eta[i]<-beta[1] + beta[2]*x[i,1] + beta[3]*x[i,2] + beta[4]*x[i,3] + beta[5]*x[i,4] + beta[6
p[i]<-exp(-exp(eta[i]))
}
#Priors
for (j in 1:6) {
beta[j] ~ dnorm(0,0.001)
```

```
}

#Prediction 1
for (i in 1:n) {
yf1[i] ~ dbin(p[i],1)
}
#Prediction 2
for (i in 1:m) {
etaf[i] <- beta[1] + beta[2]*xf[i,1] + beta[3]*xf[i,2] + beta[4]*xf[i,3] + beta[5]*xf[i,4] +
pf[i] <- exp(-exp(etaf[i]))
yf2[i] ~ dbin(pf[i],1)
}
}
```

## 6.6. Código BUGS C-Log Log

```
model
{
#Likelihood
for (i in 1:n) {
y[i] ~ dbin(p[i],1)
cloglog(p[i])<-beta[1] + beta[2]*x[i,1] + beta[3]*x[i,2] + beta[4]*x[i,3] + beta[5]*x[i,4] +
}
#Priors
for (j in 1:6) {
beta[j] ~ dnorm(0,0.001)
}

#Prediction 1
for (i in 1:n) {
yf1[i] ~ dbin(p[i],1)
}
#Prediction 2
for (i in 1:m) {
cloglog(pf[i]) <- beta[1] + beta[2]*xf[i,1] + beta[3]*xf[i,2] + beta[4]*xf[i,3] + beta[5]*xf
yf2[i] ~ dbin(pf[i],1)
}
}
```