

Aplicación del Machine Learning en la estimación del riesgo de incumplimiento en el sector bancario

Diego Alejandro Castro Llanos

Docente

Facultad de Ciencias Básicas

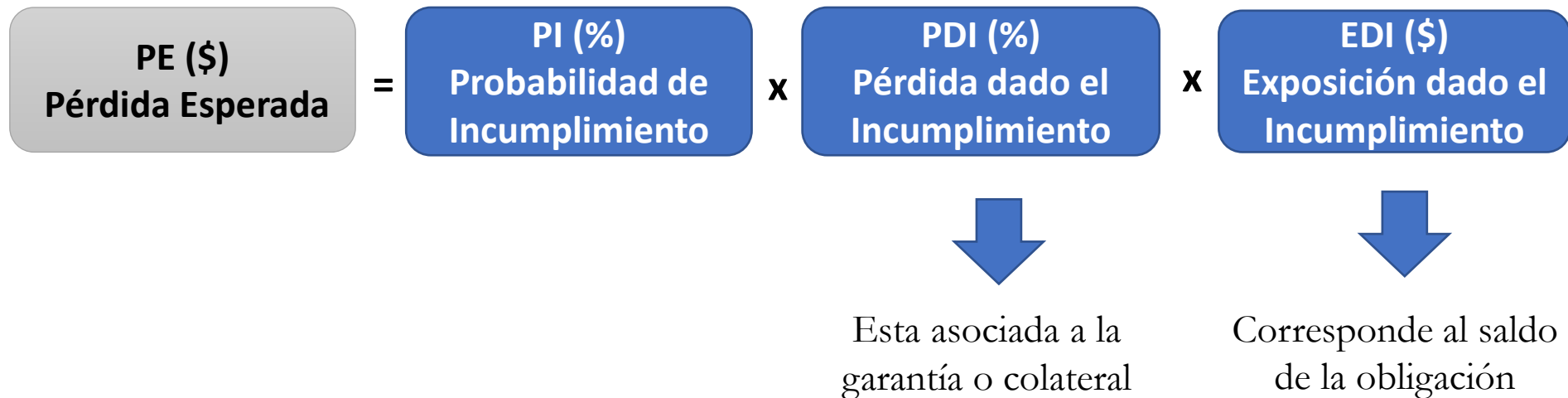
Universidad Autónoma de Occidente

La razón principal por la que existen los intermediarios financieros en una economía es captar los fondos de los agentes prestamistas o ahorradores y posteriormente dirigir los depósitos hacia los agentes deficitarios por medio de una obligación contractual. De esta manera se puede financiar proyectos en el mercado financiero e impulsar el crecimiento económico.

Debido a la **crisis financiera de 1998**, la Superintendencia Financiera de Colombia (SFC), solicitó a todas las entidades emplear modelos estadísticos que permitan estimar la probabilidad de que un cliente no cumpla con el pago de sus obligaciones durante un periodo de doce meses ($t + 12$), bajo la condición de que durante el periodo actual (t) ha sido cumplido con sus pagos. También se conoce como riesgo de *default*.

El riesgo crediticio es uno de los más importante que enfrenta el sistema financiero (SF), por lo que es necesario cuantificarlo mensualmente, monitorearlo de forma continua de acuerdo con las características actuales de la economía.

El riesgo de incumplimiento está representado por las pérdidas esperadas, que bajo norma colombiana se estiman como:



Bajo normativa internacional (NIIF 9 o IFRS 9 por sus siglas en inglés) se distinguen tres etapas:

Etapa 1: clientes cumplidos

PI estimada durante el primer año

Saldo de la obligación durante el primer año

$$PE = PI_1 * PDI_1 * EDI_1$$

Etapa 2: clientes con indicios de deterioro

PI estimada durante el segundo año

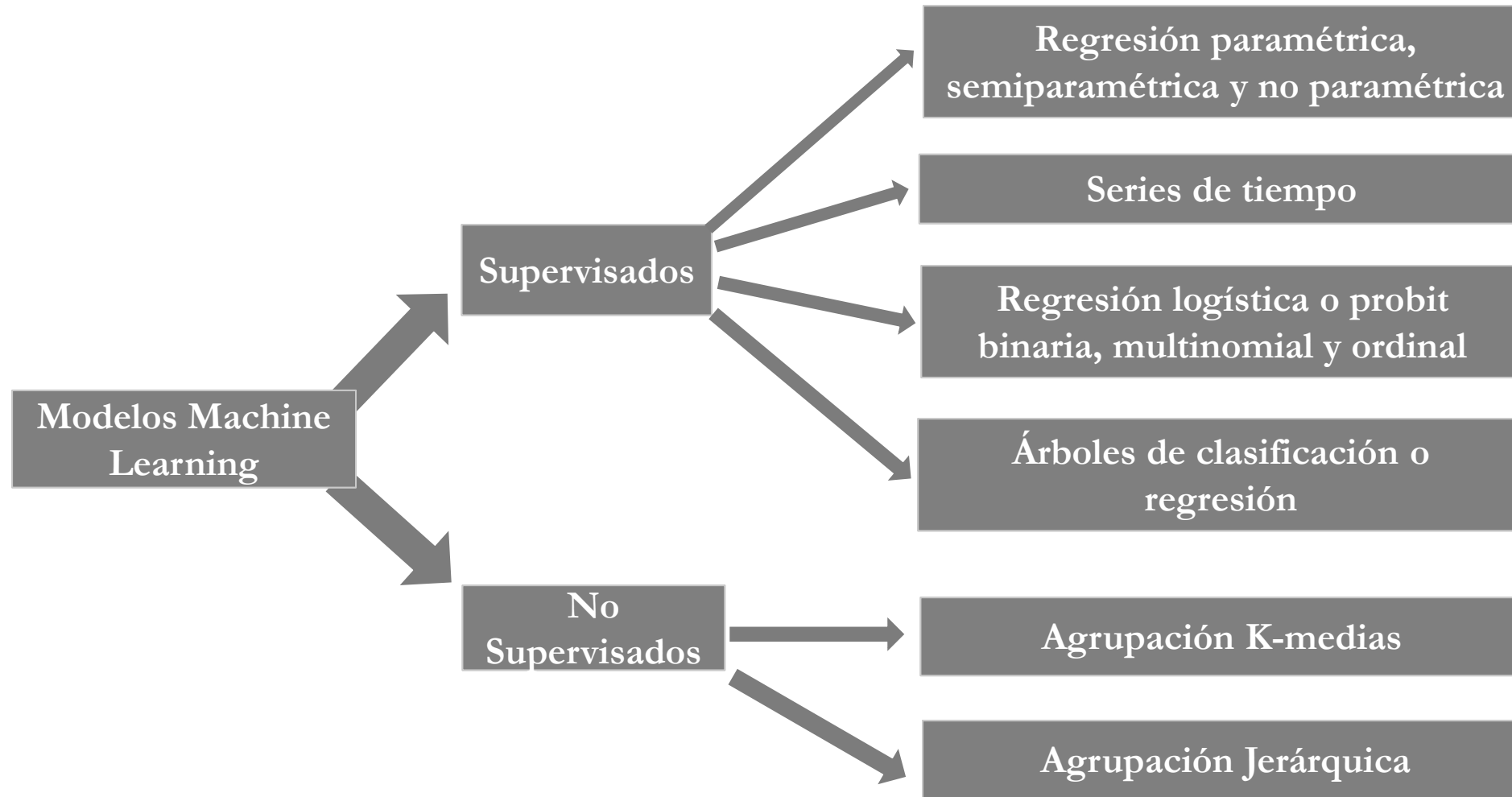
Probabilidad de supervivencia estimada durante el primer año

$$PE = PI_1 * PDI_1 * EDI_1 + \frac{PI_2 * PS_1 * PDI_2 * EDI_2}{(1 + TIR)^1} + \frac{PI_3 * PS_2 * PS_1 * PDI_3 * EDI_3}{(1 + TIR)^2} + \dots + \frac{PI_i * (\prod_{j=1}^{i-1} PS_j) * PDI_i * EDI_i}{(1 + TIR)^{i-1}}$$

En etapa 2, las pérdidas esperadas se estiman para la vida remanente del crédito. La mora es al menos de 30 días, presenta un incremento significativo de riesgo o ajuste cualitativo.

Etapa 3: clientes incumplidos

$$PE = PI_1 * PDI_1 * EDI_1 \quad ; \quad \text{donde } PI_1 = 100\%$$



En los modelos no supervisados, existen grupos: i) análisis cluster, ii) reducción de dimensionalidad.

En riesgo de crédito Comercial, la variable objetivo se puede definir como:

$$Y_i = \begin{cases} 1 & ; \text{ si el cliente es incumplido con sus obligaciones} \\ 0 & ; \text{ si el cliente es cumplido con sus obligaciones} \end{cases}$$



Se puede definir por las categorías de riesgo (A, B, C, D y E).

La definición del incumplimiento ($Y_i = 1$) depende de la norma. Específicamente:

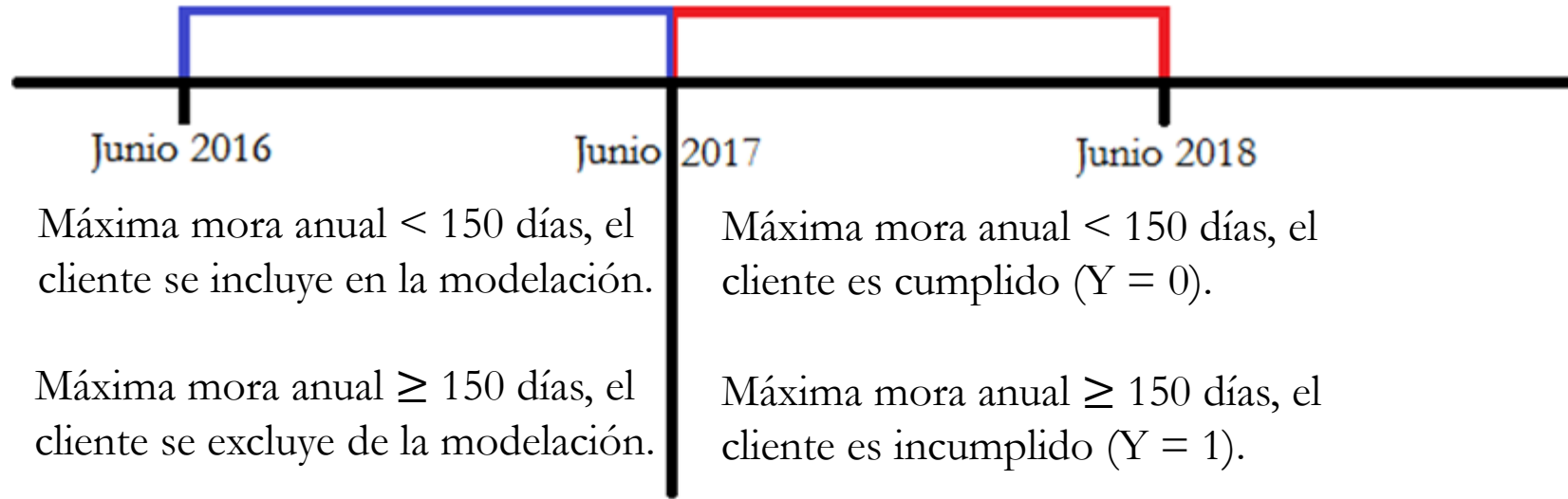
Norma	Mora (en días)
Colombiana	150
Internacional	90

Como la variable objetivo se define de manera independiente en cada uno de las normas, es necesario, estimar modelos estadísticos de manera independiente.

Variable dependiente en la estimación de los modelos de riesgo

En el portafolio Comercial, el riesgo de incumplimiento de los clientes se estima con información de los indicadores microeconómicos del balance general que entrega el cliente al intermediario financiero, comportamiento de pago y variables macroeconómicas que permitan analizar el efecto de la dinámica de la economía sobre la probabilidad de incumplimiento.

Supongamos que el balance general de un cliente corresponde al corte de junio de 2017.



$$P(\text{Incumplido } t + 12 | \text{Cumplido } t)$$

Dobson & Barnett (2008) mencionan que los Modelos Lineales Generalizados se representan mediante una combinación lineal de variables regresoras que permiten predecir una respuesta por medio de una función de enlace $g(\cdot)$, cuyo modelo se expresa de la siguiente manera:

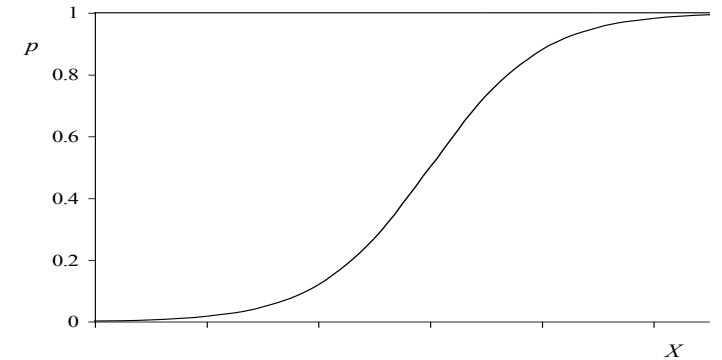
$$Y_i = E(Y_i) + \varepsilon_i; \quad E(Y_i) = \mu_i$$

$$g(E(Y_i)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad \longrightarrow \quad k \text{ variables regresoras}$$

$$g(E(Y_i)) = \alpha + X^T \beta = Z$$

Por ejemplo, la función de enlace logit se puede escribir como:

$$g(E(Y_i)) = \ln\left(\frac{p_i}{1 - p_i}\right) \quad ; \quad \text{donde} \quad p_i = \frac{1}{1 + e^{-(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}} = \Lambda(\alpha + X^T \beta)$$



Representación de los modelos de riesgo crediticio en forma reducida

En la estimación de la probabilidad de pertenecer a cada una de las categorías de riesgo (A, B, C, D y E), se pueden emplear los modelos logísticos ordinales, puesto que existe un orden jerárquico entre las categorías.

$$P(Y = A|X) = \Lambda(\alpha_A - X^T \beta) = \frac{1}{1 + e^{-(\alpha_A - X^T \beta)}} \quad ; X^T \beta = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

$$P(Y = B|X) = \Lambda(\alpha_B - X^T \beta) - \Lambda(\alpha_A - X^T \beta) = \frac{1}{1 + e^{-(\alpha_B - X^T \beta)}} - \frac{1}{1 + e^{-(\alpha_A - X^T \beta)}}$$

$$P(Y = C|X) = \Lambda(\alpha_C - X^T \beta) - \Lambda(\alpha_B - X^T \beta) = \frac{1}{1 + e^{-(\alpha_C - X^T \beta)}} - \frac{1}{1 + e^{-(\alpha_B - X^T \beta)}}$$

$$P(Y = D|X) = \Lambda(\alpha_D - X^T \beta) - \Lambda(\alpha_C - X^T \beta) = \frac{1}{1 + e^{-(\alpha_D - X^T \beta)}} - \frac{1}{1 + e^{-(\alpha_C - X^T \beta)}}$$

$$P(Y = E|X) = 1 - \Lambda(\alpha_D - X^T \beta) = \Lambda(-\alpha_D + X^T \beta) = \frac{1}{1 + e^{-(-\alpha_D + X^T \beta)}}$$

Con el signo de los parámetros estimados, solo se puede determinar si hay un aumento o disminución de la probabilidad de incumplimiento de los clientes. Para poder cuantificar el impacto en puntos porcentuales es necesario estimar **los efectos marginales**.

El efecto marginal expresa el cambio de la variable dependiente provocado por un cambio unitario en una de las variables independientes, manteniendo en *ceteris paribus* el resto de las variables regresoras.

$$\frac{\partial P(Y_i = 1)}{\partial x_i} = [\Lambda'(\alpha - x_i^T \beta)] * \beta_i$$

En caso de que se quiera expresar el cambio por cada 10 unidades, es necesario multiplicar el efecto marginal por 10.

El Accuracy Ratio (AR) o Coeficiente de Gini permite analizar la capacidad de discriminación de clientes con un buen comportamiento de pago (cumplidos) y clientes con un mal comportamiento de pago (incumplidos). Este indicador oscila entre 0% y 100%.

Altman, Haldeman & Narayanan (1977), señalan que un modelo será adecuado siempre y cuando el indicador AR sea superior o igual al 50%. Un modelo será mejor que otro entre mayor sea el indicador.

El indicador se calcula como:

$$AR = (AUC - 0,5) * 2$$

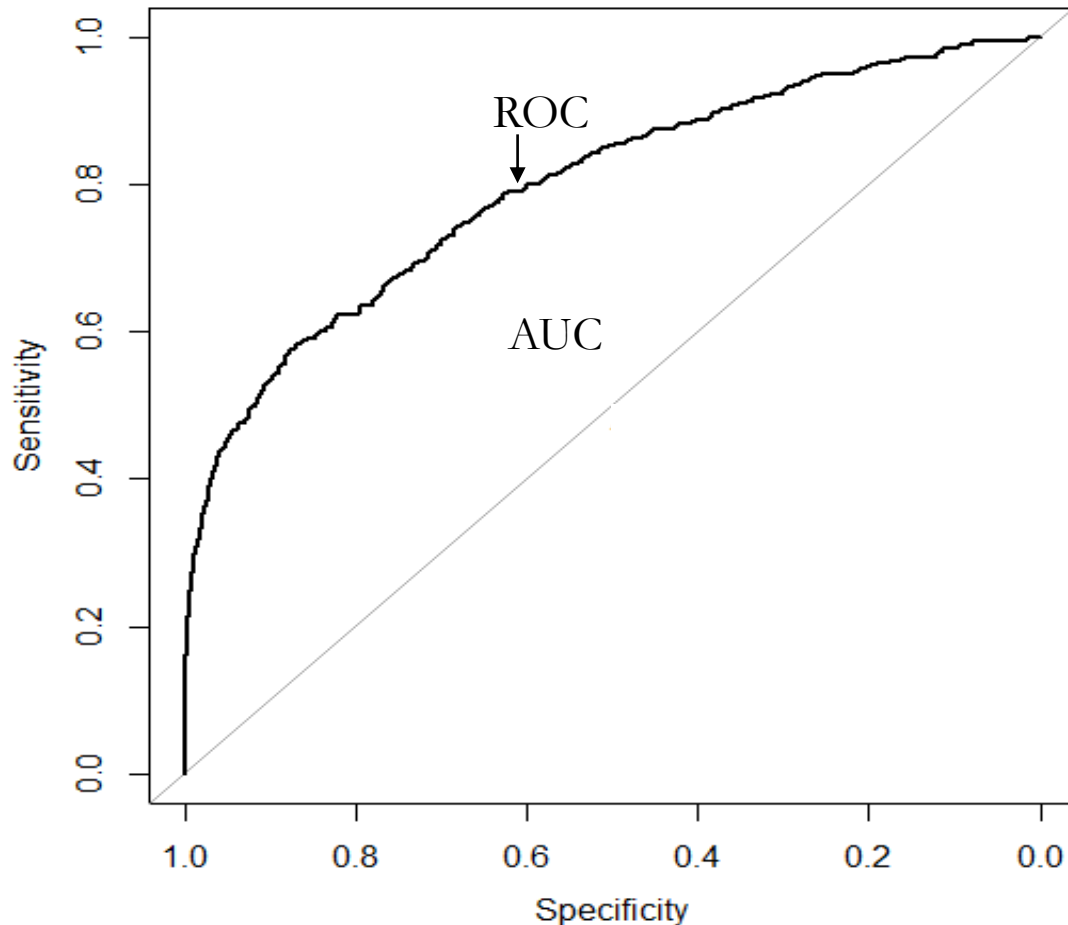
AUC es la abreviación para el área bajo la curva ROC

ROC es la abreviación para la curva característica de operación

Estimación del riesgo de incumplimiento de los clientes

Una de las medidas de bondad utilizadas en los modelos, es predecir con el modelo los valores de la variable Y de tal manera que:

- Toma el valor unitario cuando la probabilidad de ocurrencia estimada por el modelo sea mayor que un determinado valor k .
- Toma el valor cero, siempre y cuando la probabilidad de ocurrencia estimada por el modelo sea menor a un determinado valor k .



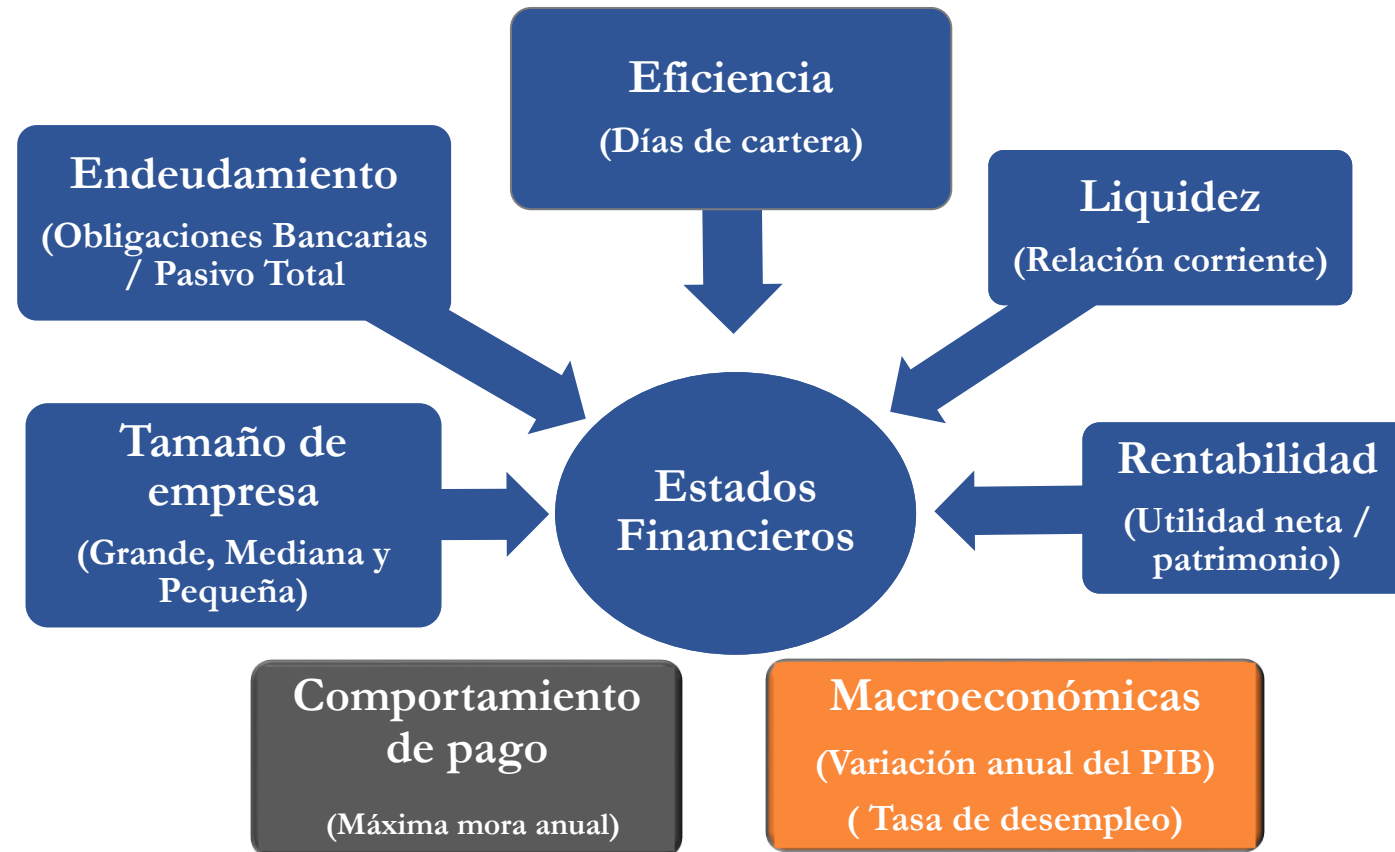
	$Y_i = 0$	$Y_i = 1$
$P_i < k$	n_{11}	n_{12}
$P_i > k$	n_{21}	n_{22}

$$\text{Especificidad} = P(P_i < 0,5 | Y_i = 0) = \frac{n_{11}}{n_{11} + n_{21}}$$

$$\text{Sensibilidad} = P(P_i > 0,5 | Y_i = 1) = \frac{n_{22}}{n_{12} + n_{22}}$$

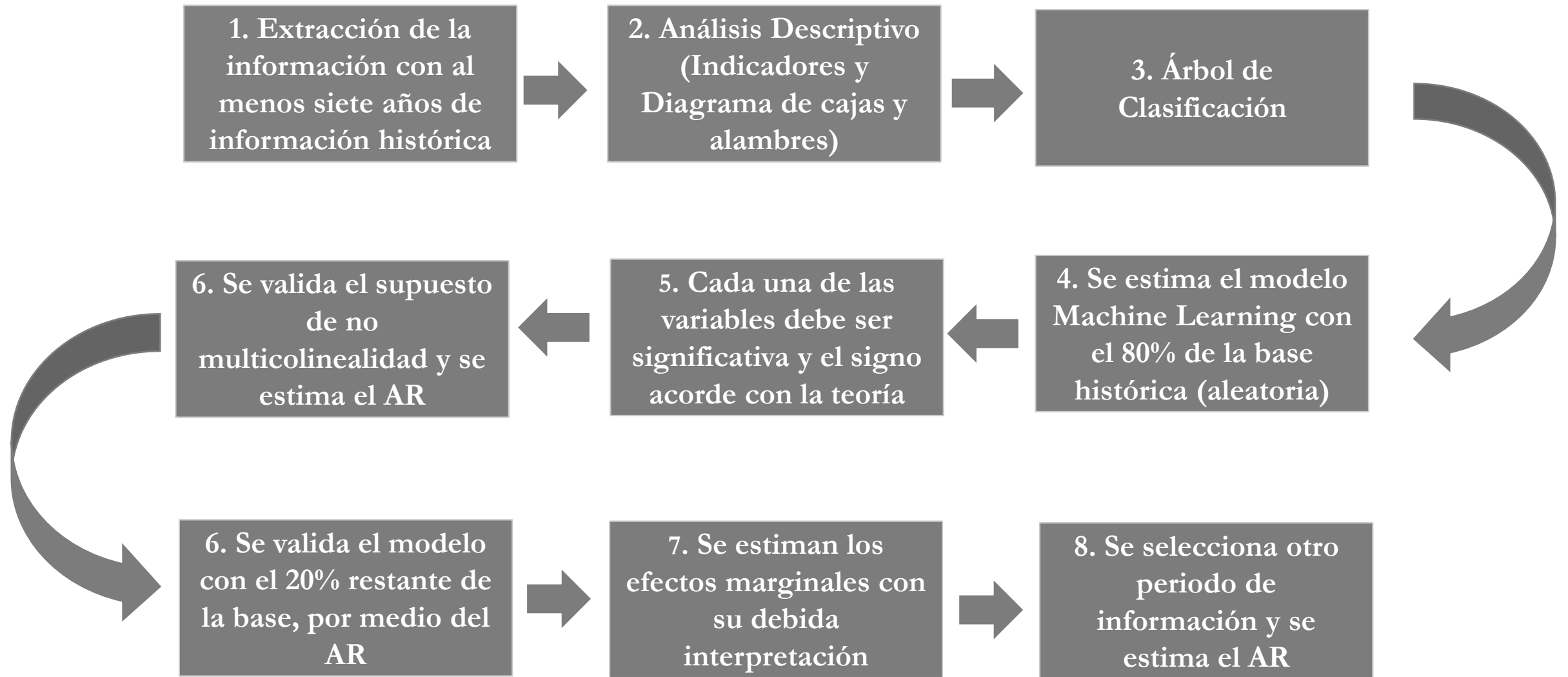
Variables microeconómicas, macroeconómicas y comportamiento de pago

En la estimación del modelo Machine Learning se tienen en cuenta variables microeconómicas de los estados financieros, comportamiento de pago de las obligaciones (mora) e indicadores macroeconómicos.



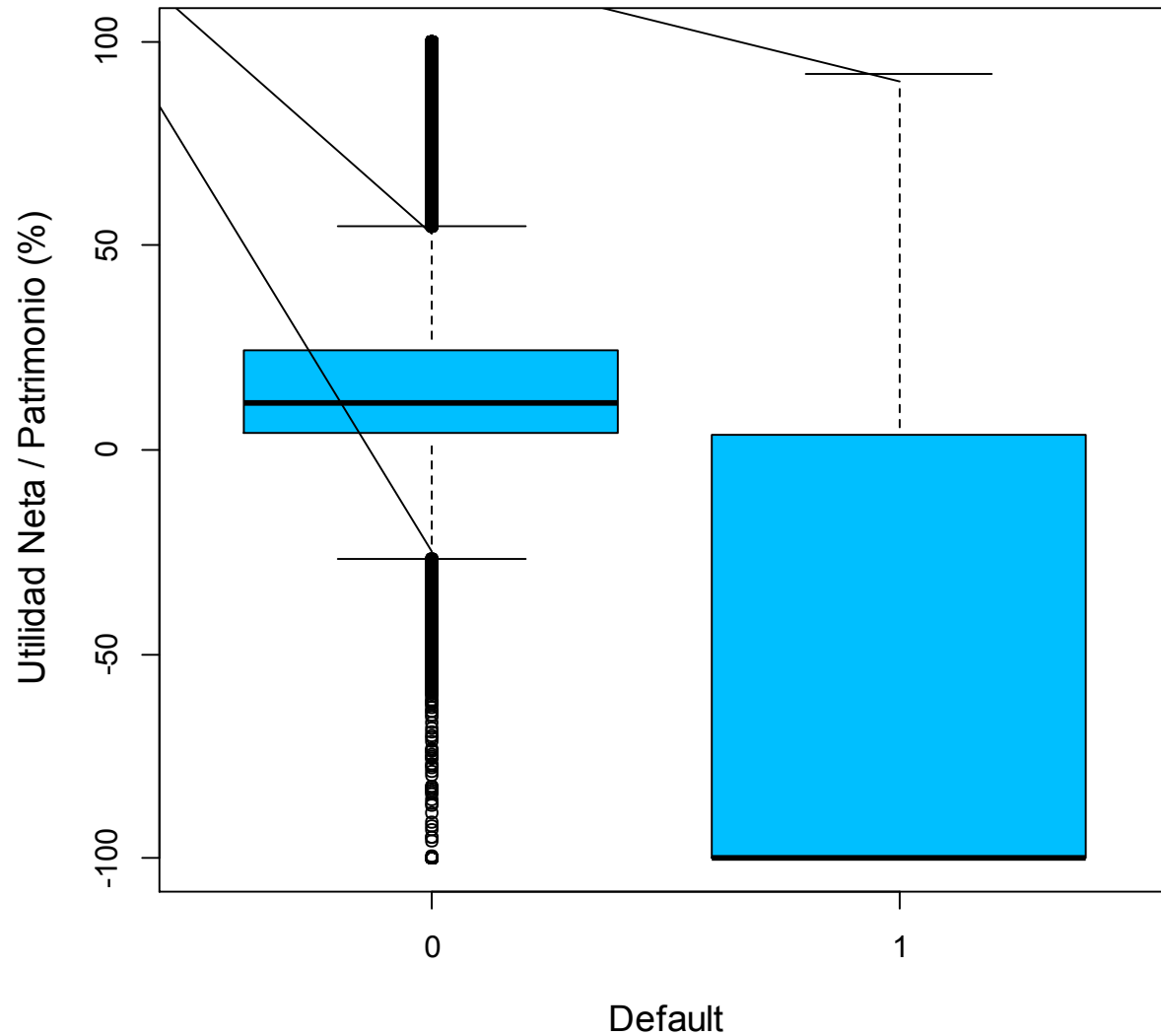
Se consideró información histórica de los clientes del sector económico Comercial entre enero de 2012 hasta diciembre de 2018.

Etapas en la estimación del modelo de riesgo de crédito

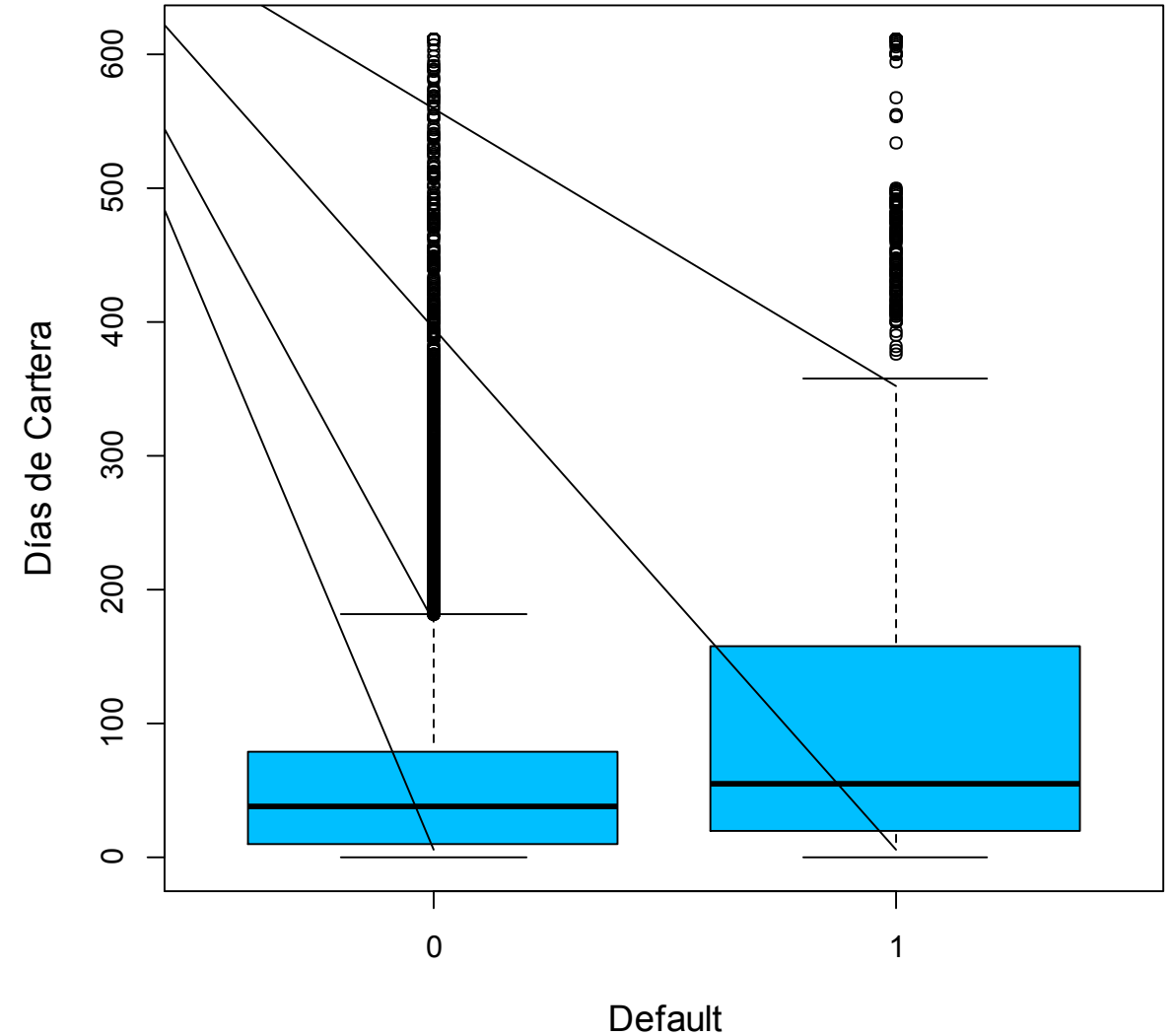


Para la estimación del modelo se cuenta con 30.141 firmas. El 80% equivale a 24.113 firmas.

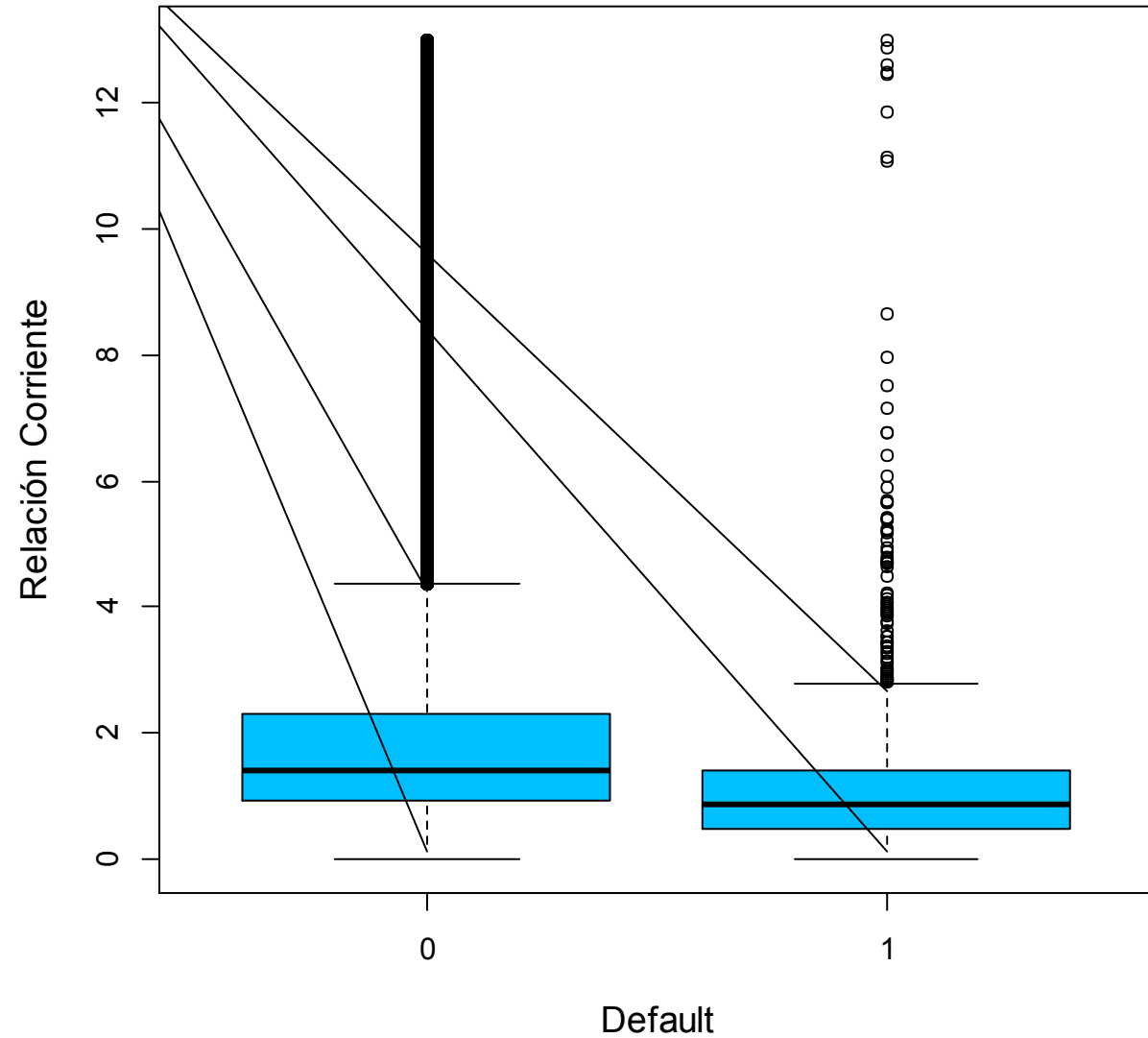
Boxplot para la Utilidad Neta / Patrimonio (%)



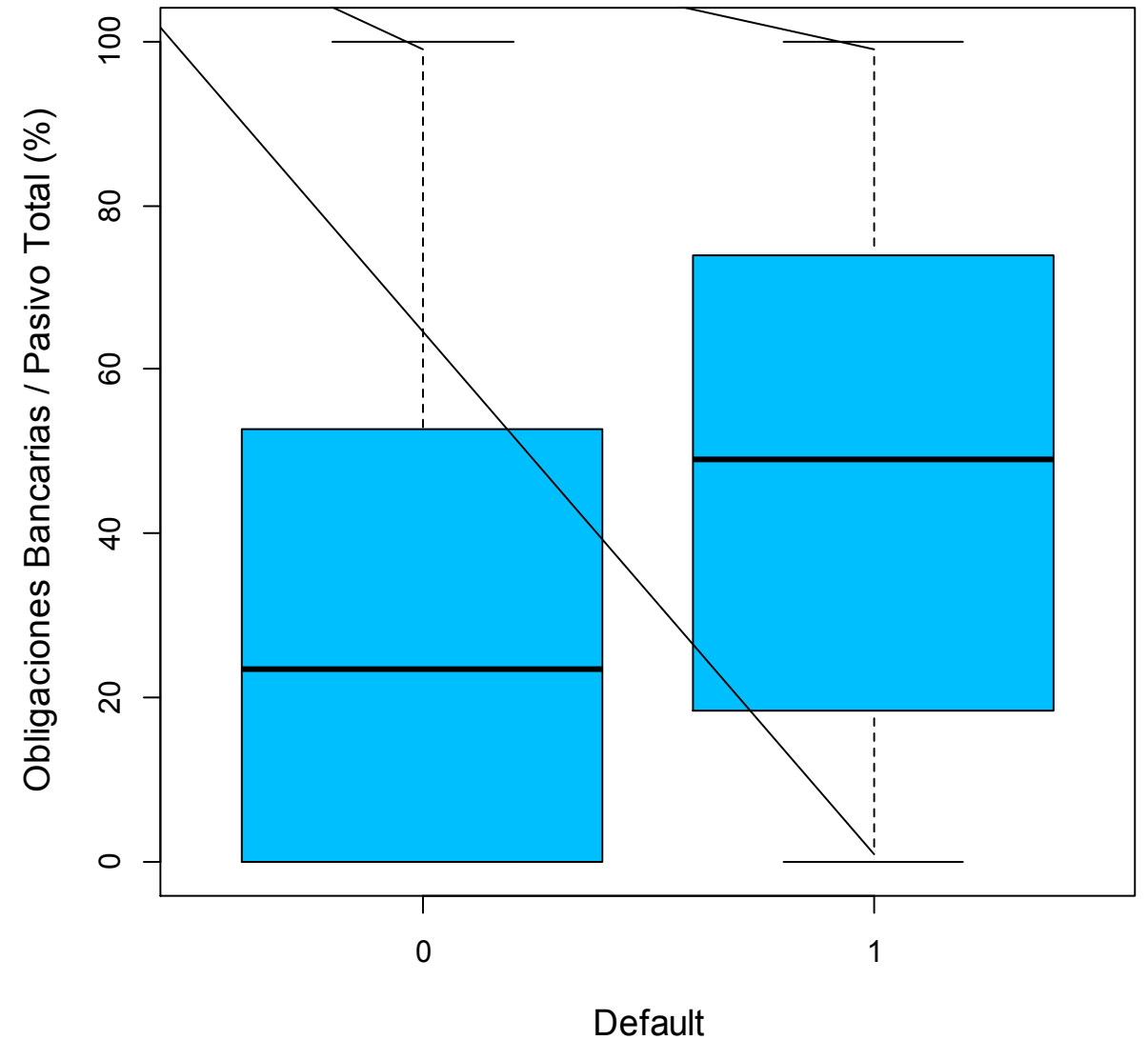
Boxplot para Días de Cartera



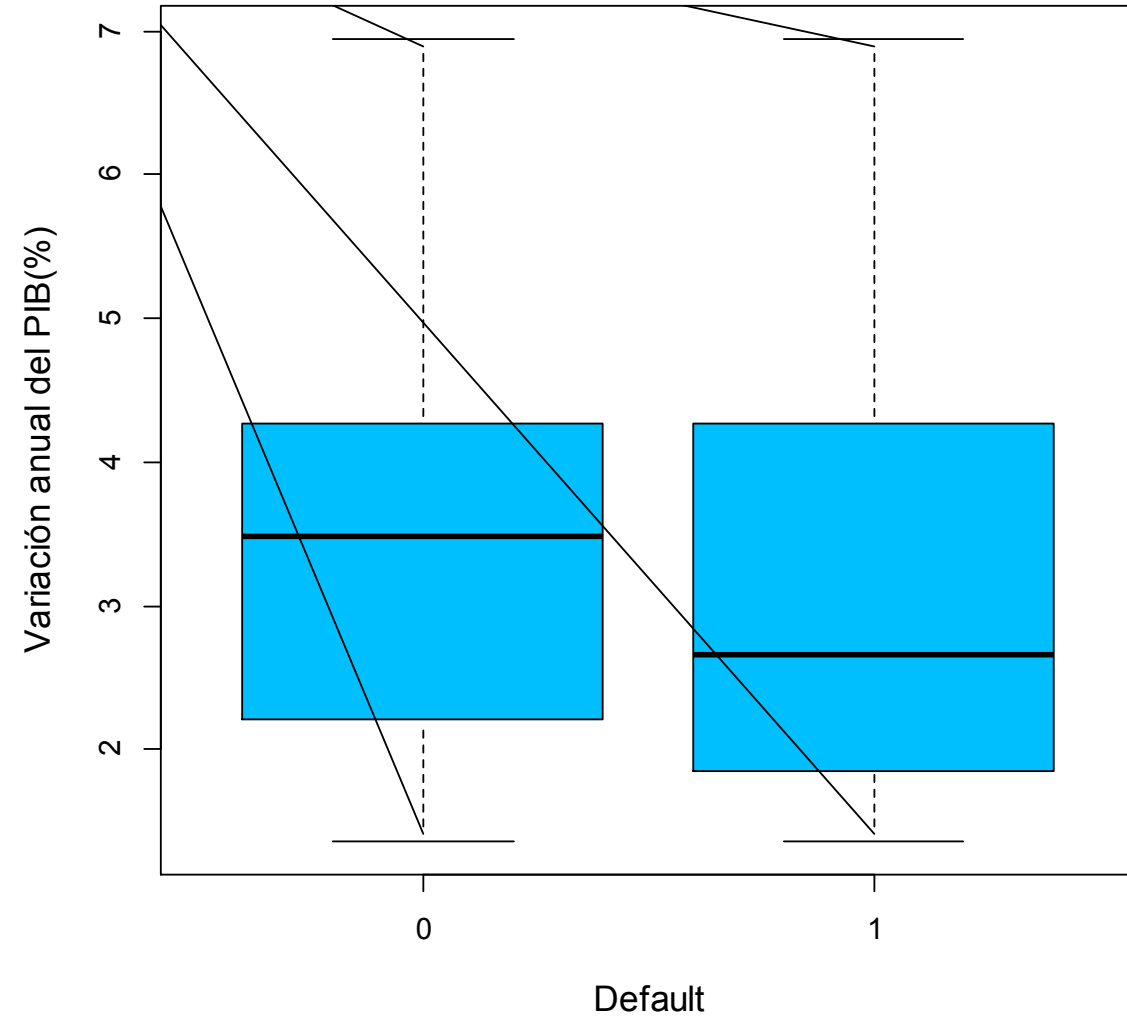
Boxplot para la Relación Corriente



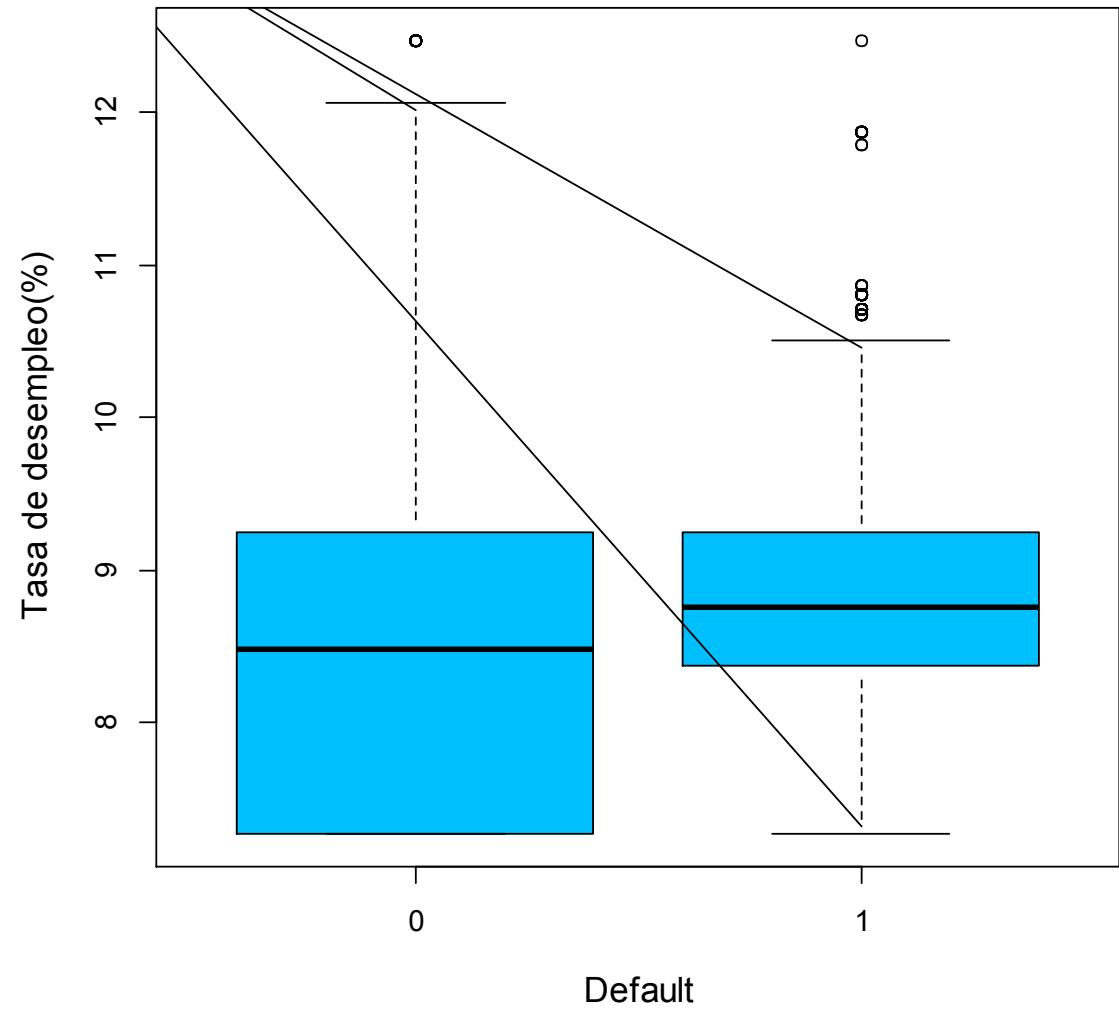
Boxplot para Obligaciones Bancarias / Pasivo Total (%)



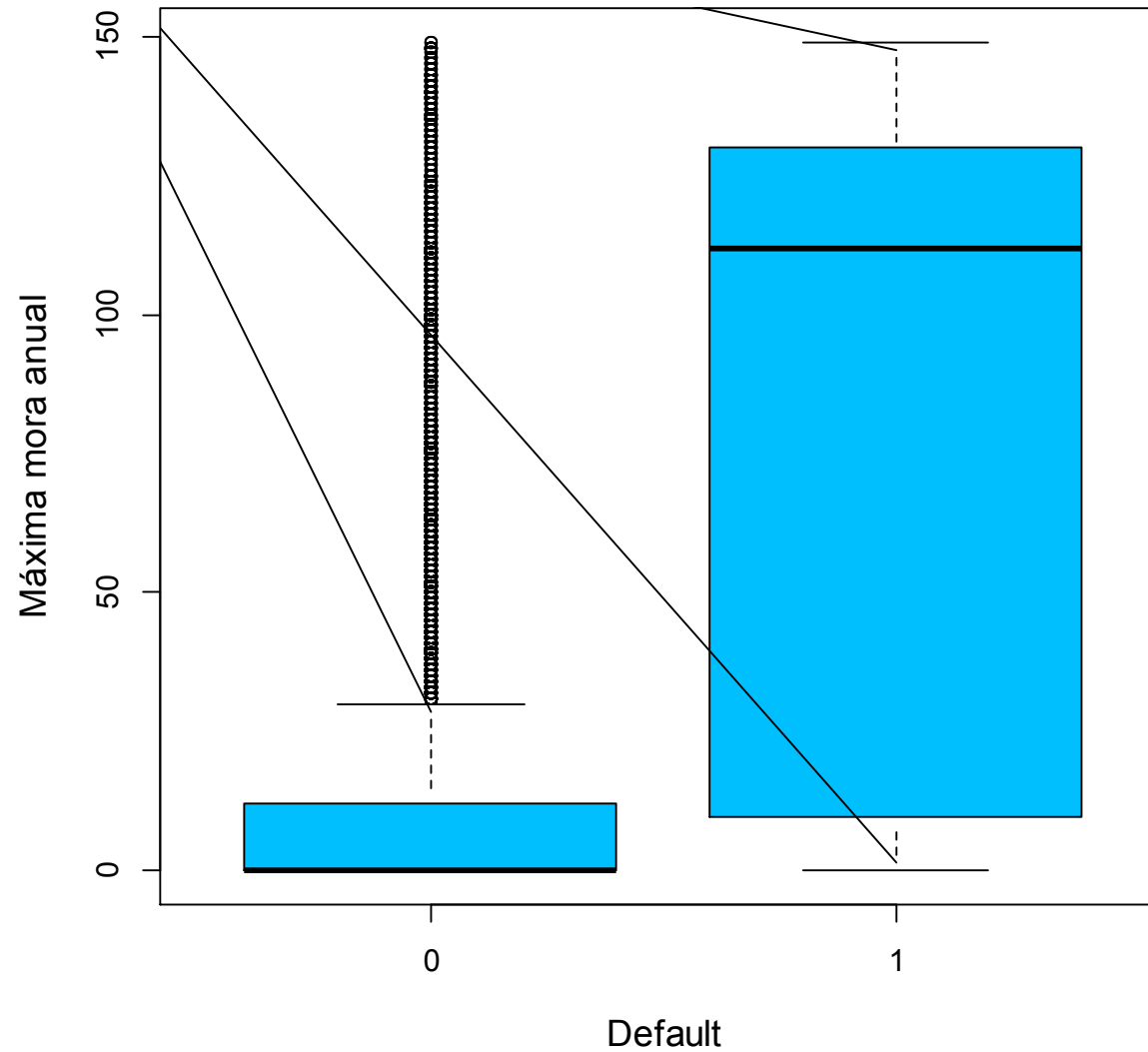
Boxplot para la Variación anual del PIB(%)



Boxplot para la Tasa de desempleo(%)



Boxplot para la Máxima mora anual



La estimación del modelo resultó ser:

Variable	Beta estimado (Error estándar)
Intercepto	-11,1686 (*) (0,5567)
Rentabilidad (x_1)	-0,0364 (*) (0,0019)
Eficiencia (x_2)	0,0028 (*) (0,0004)
Liquidez (x_3)	-0,1399 (*) (0,0304)
Endeudamiento (x_4)	0,0087 (*) (0,0016)
Mediana (x_5)	0,7299(*) (0,1993)
Pequeña (x_6)	1,1404 (*) (0,1955)
Variación anual del PIB (x_7)	-0,2455 (*) (0,0392)

Variable	Beta estimado (Error estándar)
Tasa de desempleo (x_8)	0,7509 (*) (0,0532)
Comportamiento de pago (x_9)	0,0134 (*) (0,0016)

(*) Coeficientes significativos al 1%

$$Z_i = -11,1686 - 0,0364(x_{1i}) + 0,0028(x_{2i}) - 0,1399(x_{3i}) + 0,0087(x_{4i}) + 0,7299(x_{5i}) + 1,1404(x_{6i}) - 0,2455(x_{7i}) + 0,7509(x_{8i}) + 0,0134(x_{9i})$$

$$p_i = \frac{1}{1 + e^{-Z_i}} ; \forall i$$

Validación del supuesto de no multicolinealidad

Variable	VIF
Rentabilidad	3,3998
Eficiencia	1,0238
Liquidez	1,0518
Endeudamiento	1,0290
Mediana	2,5907
Pequeña	3,2421
Variación anual del PIB	1,3766
Tasa de desempleo	1,5959
Comportamiento de pago	3,4385

Cuando se estima un modelo se parte del supuesto que las variables regresoras no presentan una relación lineal.

Uno de los métodos es el Valor de Inflación de Varianza (VIF).

No se presentan problemas de multicolinealidad, siempre y cuando el VIF sea menor a 5.

El Accuracy Ratio o Coeficiente de Gini resultó ser:

$$AR = (AUC - 0,5) * 2$$

$$AR = (0,888 - 0,5) * 2 = 0,776$$

$$AR = 77,6\% > 50\%$$

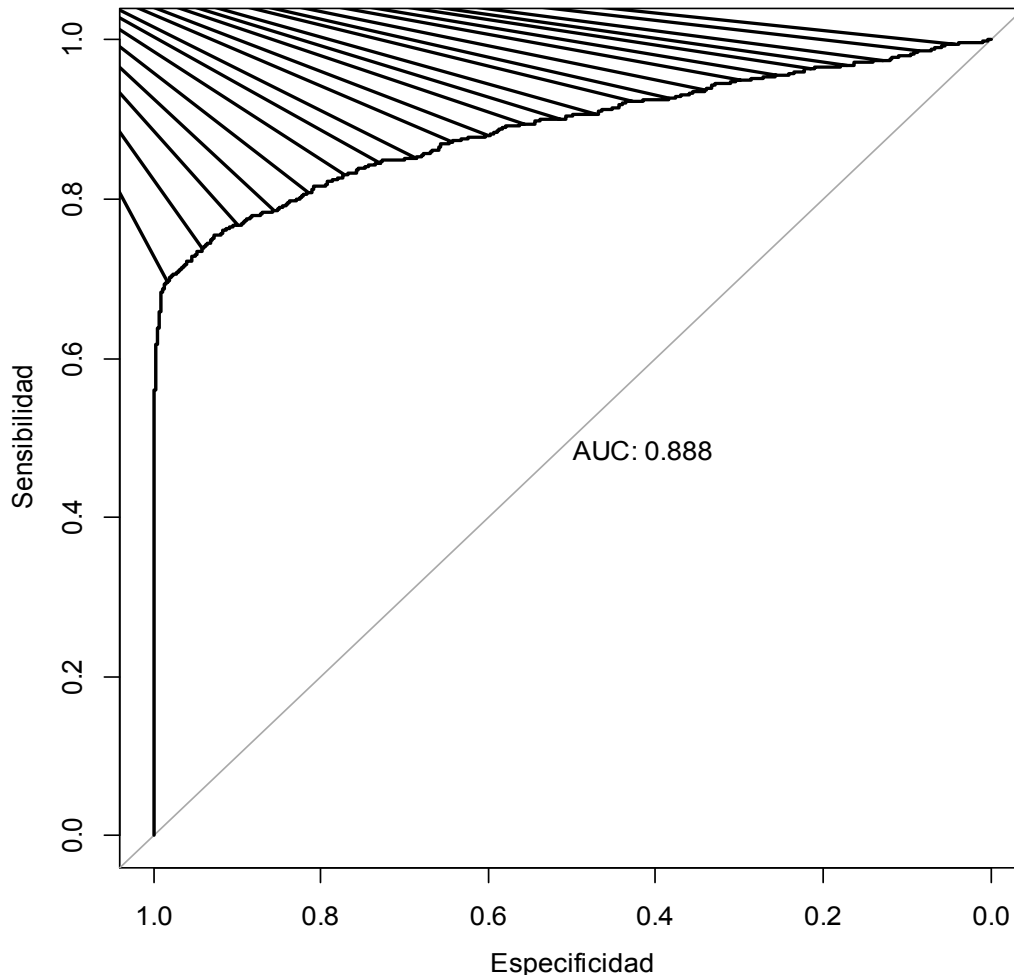
$$Sensibilidad = 0,6212 = 62,12\%$$

De cada 100 clientes incumplidos con el pago de las obligaciones, se espera que el modelo logre captar 62 incumplidos.

$$Especificidad = 0,9959 = 99,59\%$$

De cada 1000 clientes cumplidos con el pago de las obligaciones, se espera que el modelo logre captar 996 cumplidos

Gráfico area bajo la curva ROC



Validación del modelo con el 20% restante de la base de modelación

Con el 20% restante de la base de modelación que corresponde a 6.028 firmas se estimó la probabilidad de incumplimiento del cliente a partir de la ecuación:

$$Z_i = -11,1686 - 0,0364(x_{1i}) + 0,0028(x_{2i}) - 0,1399(x_{3i}) + 0,0087(x_{4i}) + 0,7299(x_{5i}) + 1,1404(x_{6i}) - 0,2455(x_{7i}) + 0,7509(x_{8i}) + 0,0134(x_{9i})$$

$$p_i = \frac{1}{1 + e^{-Z_i}}$$

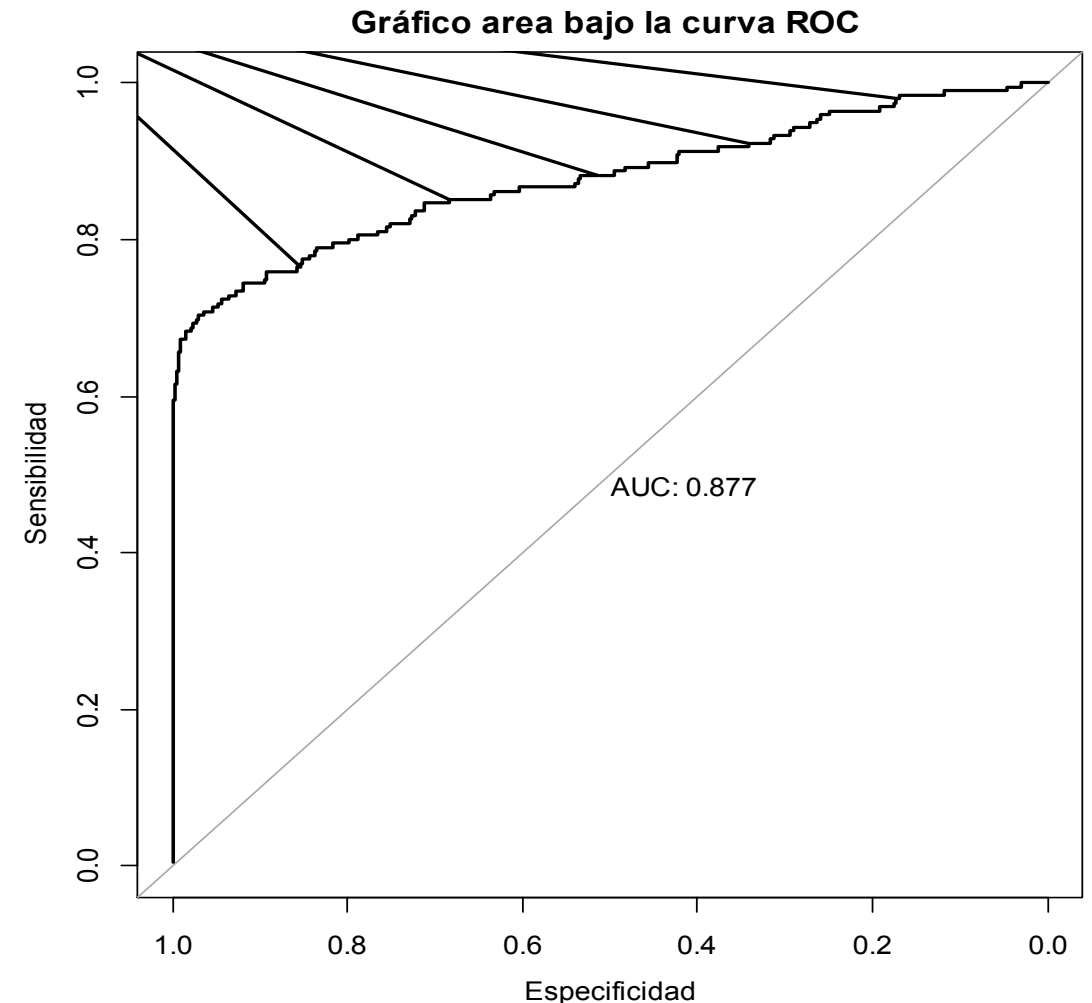
$$AR = (AUC - 0,5) * 2$$

$$AR = (0,877 - 0,5) * 2 = 0,754$$

$$AR = 75,4\%$$

$$Sensibilidad = 0,6154 = 61,54\%$$

$$Especificidad = 0,9967 = 99,67\%$$



Variable	Efecto marginal
Rentabilidad	-0,063
Eficiencia	0,005
Liquidez	-0,242
Endeudamiento	0,015
Mediana	1,263
Pequeña	1,974
Variación anual del PIB	-0,425
Tasa de desempleo	1,30
Comportamiento	0,023

pp = puntos porcentuales

Por cada 10 puntos porcentuales que aumente la rentabilidad de la firma, se espera que la probabilidad de incumplimiento (PI) disminuye en 0,63 pp, estando las demás variables constantes.

Por cada 100 puntos porcentuales que aumente la variable de eficiencia, se espera que la PI aumente en 0,50 pp, estando las demás variables constantes.

Por cada 10 puntos porcentuales que aumente el nivel de endeudamiento de la firma, se espera que la PI aumente en 0,15 puntos porcentuales, estando las demás en *ceteris paribus*.

Variable	Efecto marginal
Rentabilidad	-0,063
Eficiencia	0,005
Liquidez	-0,242
Endeudamiento	0,015
Mediana	1,263
Pequeña	1,974
Variación anual del PIB	-0,425
Tasa de desempleo	1,30
Comportamiento	0,023

pp = puntos porcentuales

Si una empresa es mediana, se espera que la PI aumente 1,26 pp respecto a las empresas grandes, siempre y cuando las demás variables sean constantes.

Si una empresa es pequeña, se espera que la PI aumente 1,97 pp respecto a las empresas grandes, siempre y cuando las demás variables sean constantes.

Por cada punto porcentual que aumente la variación anual del PIB, se espera que la PI disminuya en promedio 0,43 pp, estando las demás variables en *ceteris paribus*.

Variable	Efecto marginal
Rentabilidad	-0,063
Eficiencia	0,005
Liquidez	-0,242
Endeudamiento	0,015
Mediana	1,263
Pequeña	1,974
Variación anual del PIB	-0,425
Tasa de desempleo	1,30
Comportamiento	0,023

pp = puntos porcentuales

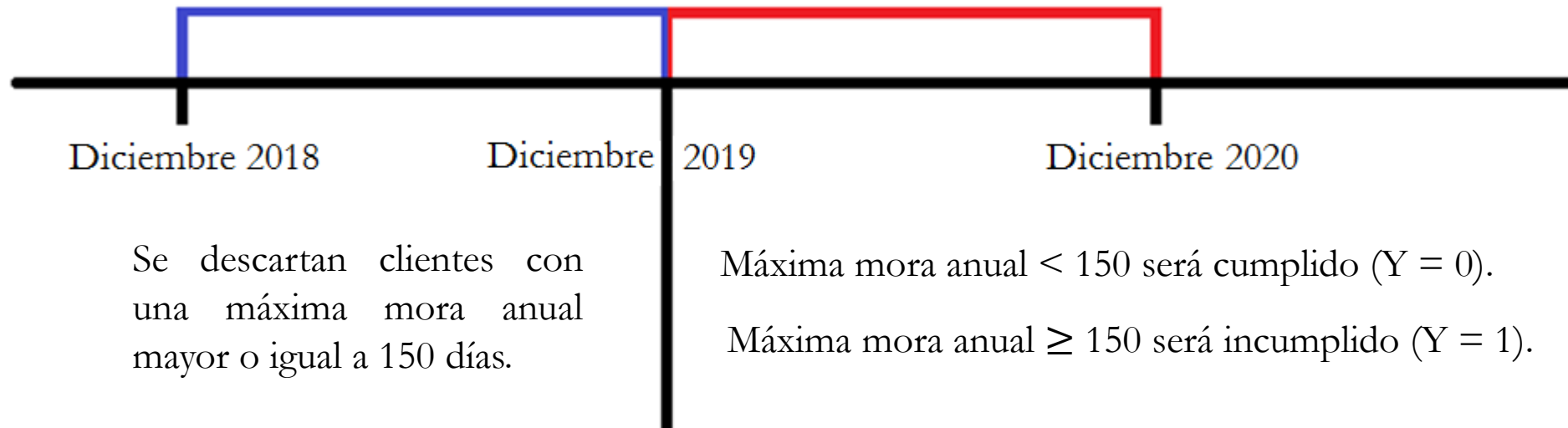
Por cada punto porcentual que se incremente la tasa de desempleo, se espera que la PI se incrementé 1,3 pp, estando las demás variables constantes.

Por cada 10 días de mora que se incremente la variable de comportamiento de pago, se espera que la PI se incremente 2,3 pp estando las demás variables en *ceteris paribus*.

$$0,023 * 10 = 2,3 \text{ pp}$$

Backtesting con información externa de la base de modelación y validación

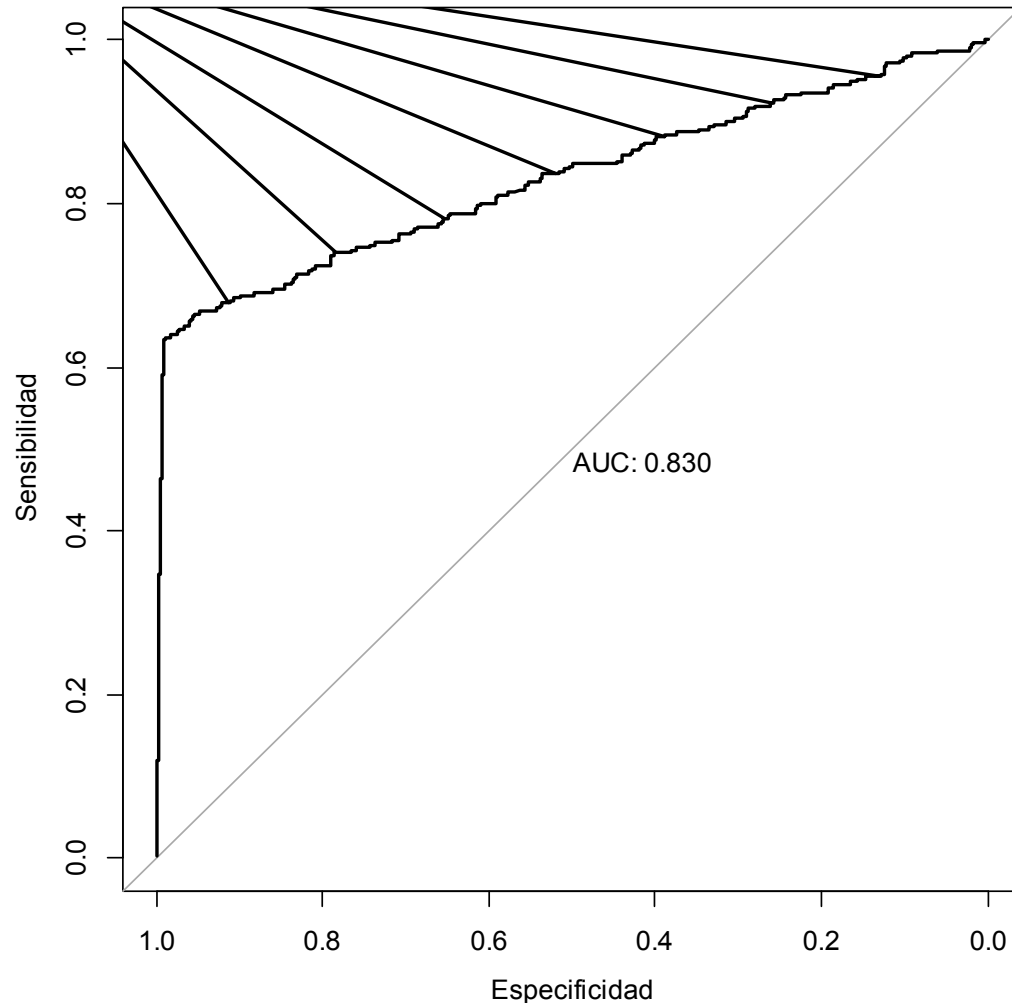
Se considera todos los clientes cumplidos al corte de diciembre de 2019 y se evalúa su comportamiento de pago durante los doce meses siguientes (enero 2020 – diciembre 2020).



Base	Conteo	Número Default	Probabilidad Real
Entrenamiento	24.113	829	3,4338%
Modelación	6.028	193	3,2017%
Backtesting	7.933	311	3,9203%

Backtesting con información externa de la base de modelación y validación

Se considera todos los clientes cumplidos al corte de diciembre de 2019 y se evalúa su comportamiento durante los doce meses siguientes (enero 2020 – diciembre 2020).



$$Z_i = -11,1686 - 0,0364(x_{1i}) + 0,0028(x_{2i}) - 0,1399(x_{3i}) + 0,0087(x_{4i}) + 0,7299(x_{5i}) + 1,1404(x_{6i}) - 0,2455(x_{7i}) + 0,7509(x_{8i}) + 0,0134(x_{9i})$$

$$p_i = \frac{1}{1 + e^{-Z_i}} \quad AR = (0,83 - 0,5) * 2 = 0,66$$

$$AR = 66\%$$

$$Sensibilidad = 0,627 = 62,7\%$$

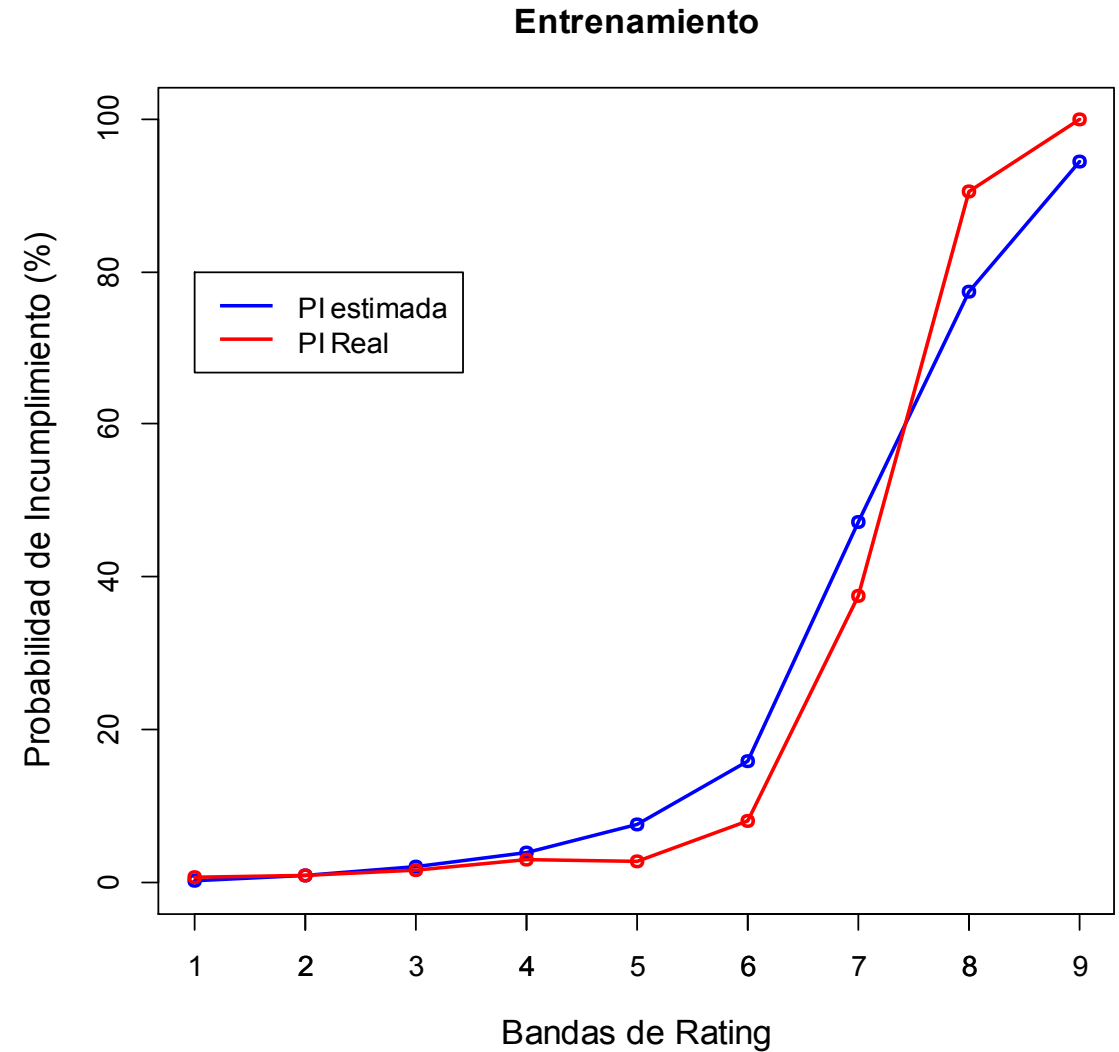
$$Especificidad = 0,9919 = 99,19\%$$

Comparativo PI estimada vs real de la base de modelación

Se analiza el comportamiento de la probabilidad de incumplimiento estimada vs real con información de la base de modelación. De un total de 24.113 clientes cumplidos, al cabo de un año 829 resultaron incumplidos.

Banda	Intervalo (%)	PI estimada (%)	PI real (%)	% de clientes
1	(0 – 0,6]	0,29	0,68	47,11
2	(0,6 - 1,5]	0,96	0,94	27,45
3	(1,5 - 3]	2,07	1,68	14,08
4	(3 – 6]	4,07	3,08	5,12
5	(6 – 10]	7,68	2,82	1,77
6	(10 – 30]	16,02	8,07	1,34
7	(30 – 60]	47,17	37,55	1,08
8	(60 – 90]	77,41	90,49	1,52
9	(90 – 100]	94,30	100	0,52
Total		3,40	3,43	

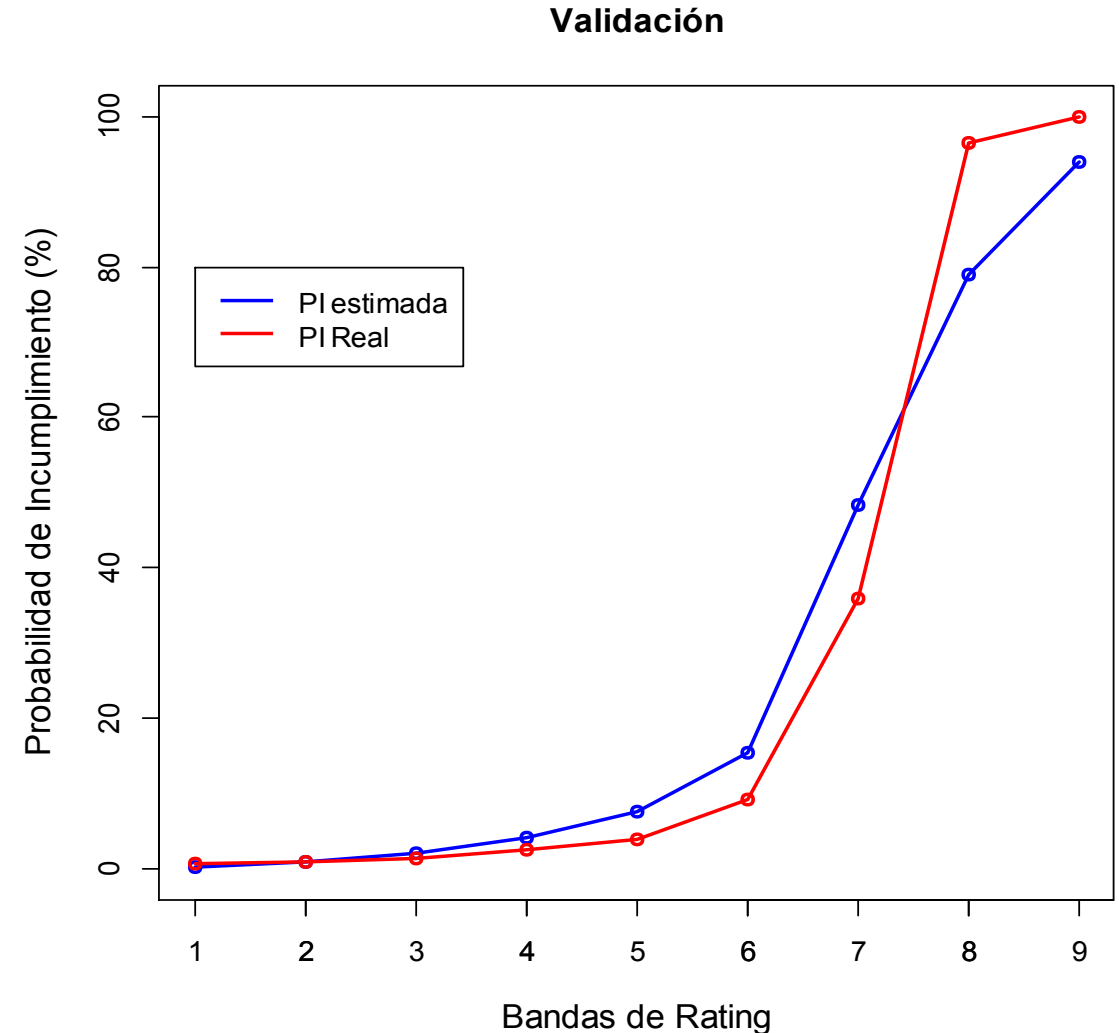
AR = 77,6%



Se analiza el comportamiento de la probabilidad de incumplimiento estimada vs real con información de la base de validación. De un total de 6.028 clientes cumplidos, al cabo de un año 193 resultaron incumplidos.

Banda	Intervalo (%)	PI estimada (%)	PI real (%)	% de clientes
1	(0 – 0,6]	0,29	0,76	45,69
2	(0,6 - 1,5]	0,97	0,92	28,90
3	(1,5 - 3]	2,07	1,50	14,41
4	(3 – 6]	4,18	2,48	5,34
5	(6 – 10]	7,70	3,92	1,69
6	(10 – 30]	15,53	9,21	1,26
7	(30 – 60]	48,36	35,84	0,88
8	(60 – 90]	78,99	96,55	1,44
9	(90 – 100]	93,90	100	0,38
Total		3,18	3,23	

AR = 75,4%

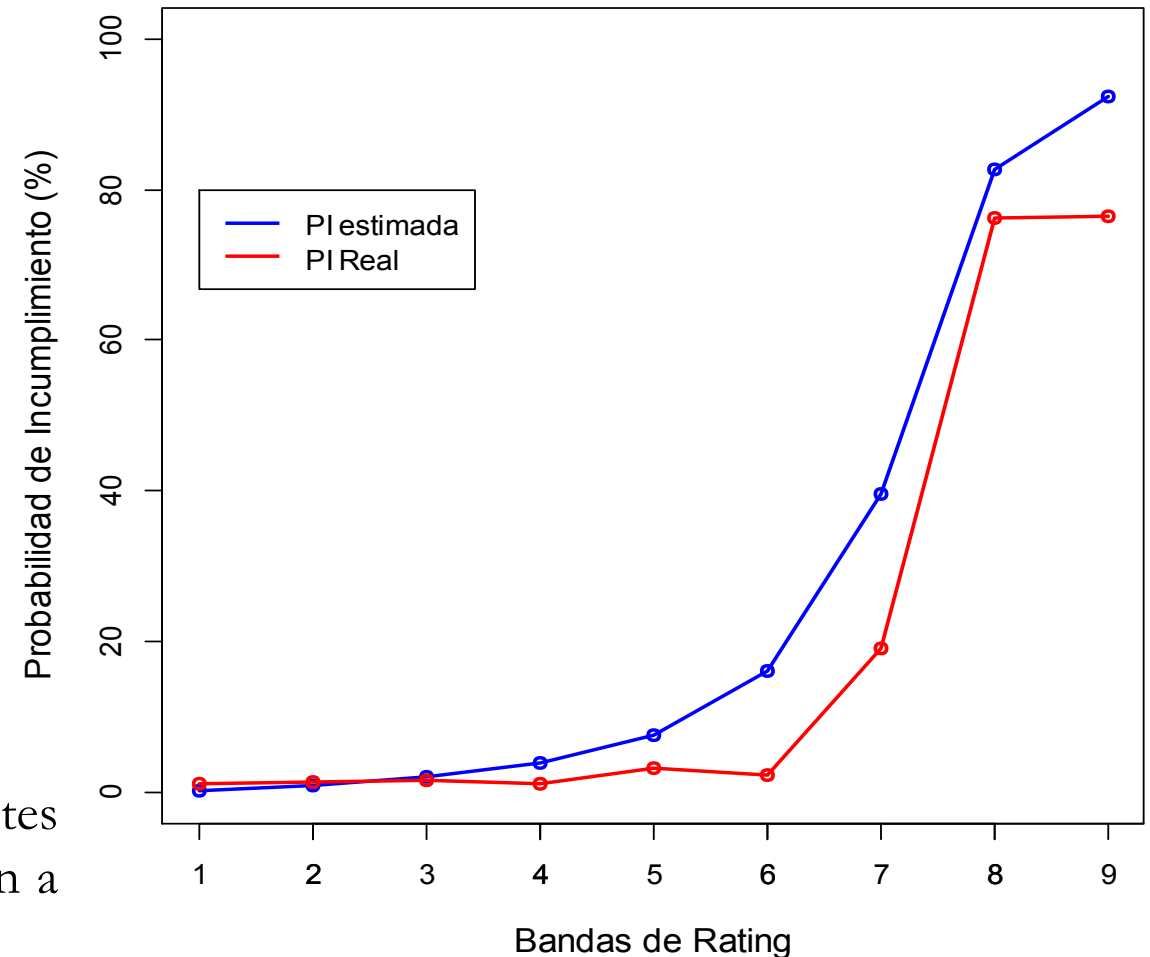


Se analiza el comportamiento de la probabilidad de incumplimiento estimada vs real con información externa. De un total de 7.933 clientes cumplidos, al cabo de un año 311 resultaron incumplidos.

Backtesting

Banda	Intervalo (%)	PI estimada (%)	PI real (%)	% de clientes
1	(0 – 0,6]	0,33	1,20	21,00
2	(0,6 - 1,5]	1,03	1,35	32,77
3	(1,5 - 3]	2,10	1,68	27,81
4	(3 – 6]	3,92	1,30	10,64
5	(6 – 10]	7,63	3,35	2,64
6	(10 – 30]	16,24	2,33	1,63
7	(30 – 60]	39,66	19,23	0,33
8	(60 – 90]	82,65	76,24	2,55
9	(90 – 100]	92,23	76,47	0,64
Total		4,70	3,92	

$AR = 66\%$ A causa del covid 19, el porcentaje de clientes incumplidos aumentó en un 14,28% en relación a la base de modelación.



El comparativo de la sensibilidad, especificidad e indicador AR son:

	Sensibilidad	Especificidad	AR
Entrenamiento	62,12%	99,59%	77,6%
Validación	61,54%	99,67%	75,4%
Backtesting	62,7%	99,19%	66%

Se explicará como se estiman los modelos que permiten predecir la probabilidad de incumplimiento para cada uno de los años de la vida remanente del crédito, teniendo en cuenta la proyección de variables macroeconómicas en tres diferentes escenarios Base, Favorable y Desfavorable