



ntic
master
School


UNIVERSIDAD
COMPLUTENSE
DE MADRID



Riesgo de Crédito y modelos de Scoring

Resumen.

En este tema analizaremos los modelos predictivos en la "gestión de riesgos":

- **Scoring de Riesgo de Crédito;** 
- **Cálculo de las Primas de Riesgo;**
- **Modelos para la detección de Fraude;**
- **Modelos de fuga de clientes.**

Repasaremos los métodos para la estimación de modelos de probabilidad; y aspectos fundamentales a tener en cuenta en este tipo de análisis como son el sesgo de selección muestral, la tramificación de variables continuas y los métodos de selección de variables explicativas.



Bibliografía básica:

- Anderson, R(2007) *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation* . Oxford University Press
- Mays,E and Niall Lynas (2011) *Credit Scoring for Risk Managers: The Handbook for Lenders*.Createspace (ISBN13: 9781450578967)
- Siddiqi, N. (2006): *Credit Risk Scorecards. Developing and implementing Intelligent Credit Scoring*. J Wiley & Sons
- Bolder, J. D. (2019). *Credit-Risk Modelling: Theoretical Foundations, Diagnostic Tools, Practical Examples, and Numerical Recipes in Python*. Springer
- Trueck, S, & Rachev, Svetlozar (2009): *Rating Based Modeling of Credit Risk. Theory and Application of Migration Matrices*. Elsevier
- Wooldridge, J.M. (2019) . *Introductory Econometrics. A modern Approach*. 7a Edicion Cengage Learning. (Cap 17)
- Géron, A. (2022): *Hands-On Machine Learning with scikit-learn, Keras and TensorFlow. Concepts, Tools and Techniques to Build Intelligent Systems*. Ed. O'Eilly

¿Qué es “Riesgo”?

Concepto de Riesgo: Asociado al concepto de **incertidumbre** sobre el **acontecimiento de un siniestro**, de que se **materialice una amenaza**, al peligro de que sucedan **pérdidas, deterioros o perjuicios o efectos adversos a nuestros intereses**

Distintos **tipos de riesgos**

- Riesgos **accidentes laborales**
- Riesgos por **desastres naturales** (terremotos, volcanes, inundaciones, etc),
- **Riesgos biológicos** (infecciones, epidemias),
- Riesgos **económicos**

**Hablaremos de riesgos sí,
... pero Financieros**

Gestión del riesgo

Identificación de amenazas y vulnerabilidades (exposición al riesgo).

- La **amenaza** o peligro latente es aquél fenómeno o suceso que de acontecer podría causarnos un perjuicio (cuando se materializa se habla de **sinistro**). Detectada una posible amenaza, el problema es que tenemos incertidumbre sobre si acontecerá o no acontecerá
- La **vulnerabilidad** es la pérdida esperada o el perjuicio esperado en caso de que se **materialice la amenaza**. Cuanto mayor es la vulnerabilidad ante una amenaza o peligro latente, mayor es el riesgo, mayores serán las pérdidas o deterioros que podría sufrir, y cuanto más factible es el perjuicio o daño mayor es el peligro

Gestión del riesgo

- **Planes de previsión:** **anticipar** la ocurrencia del siniestro Intentar predecir, cuantificar la probabilidad de que ocurra
- **Planes de prevención:** se pretende **reducir la probabilidad** de que suceda el siniestro – de riesgos laborales, prevención de incendios –
- **Planes para reducir o mitigar los efectos de un desastre** –cuando sucede el siniestro, intentar que éste cause el menor daño posibles ej. planes de evacuación, simulacros, cobertura de riesgos financieros ... se pretende **reducir la vulnerabilidad**
- y **Planes de recuperación** en caso de desastre – por ejemplo la declaración de “zona catastrófica” para disponer de fondos extraordinarios para financiar la recuperación de las zonas que hayan sufrido inundaciones, terremotos, etc- otro ejemplo es la **contratación de pólizas de seguros** de vida, del hogar, de vehículos .

Riesgos en sentido económico

PERDIDAS económicas (reducción de rentas o ingresos, reducción de riqueza, reducción de beneficios o que estos se vuelvan negativos)

Gestión del Riesgo:

- **¿Cuál es la probabilidad de que suceda un siniestro?**
- **¿y si al final sucede, cuales serán mis pérdidas?**
- **¿puedo reducir la amenaza?**
- **¿puedo reducir mi vulnerabilidad o exposición al riesgo?**
- **¿puedo cubrirme o asegurar el riesgo?**

Tres tipos de riesgos económicos según su procedencia:

- **Riesgos de la Unidad Estratégica de Negocio** (propios de la empresa). Ejemplo: riesgo por el lanzamiento de un nuevo producto, o una nueva inversión, ¿saldrá bien, será rentable?
- **Riesgos del Entorno Económico**. Hacen referencia a peligros procedentes de **cambios en el entorno o contexto económico y político**, ¿qué efectos tendrá el corona virus sobre la Economía Mundial? ¿qué pasará con la economía española si se declara un segundo confinamiento o viene otra nueva variante letal?, ¿qué pasará si el BCE y la UE no nos inyectan más dinero? ¿qué pasará si la inflación europea no se relaja y los tipos de interés continúan tan elevados? ¿qué pasará si se independiza Cataluña? ¿y que pasaría se se quiere dar la vuelta al Brexit y el Reino unido desea volver a la UE? ¿y qué pasará si Puting declara la guerra a la OTAN? ¿Conseguirá Pedro Sánchez terminar la legislatura? ¿Qué efectos tendrá el fanatismo islámico? ¿habrá una tercera guerra mundial?, ¿es el covid parte de una guerra biológica entre las grandes potencias mundiales?
- **Riesgos Financieros**: se derivan de las pérdidas originadas de los activos financieros

¿Qué son y para que sirven los Activos Financieros?

Activos Financieros: Derecho de cobro de una determinada cantidad de dinero en un determinado plazo (para el que la emite es un **pasivo financiero**, una deuda, una obligación de pago de una determinada cantidad de dinero)

Sirven **para mantener la riqueza** (forma de ahorro) **pero movilizandolos flujos financieros:** para que el ahorro fluya hacia la inversión

Los diferentes **tipos de activos financieros** se diferencian por sus características de:

- **Rentabilidad**

- **Liquidez**

- **Riesgo**

- a mayor liquidez, menor rentabilidad y menor riesgo
- a mayor rentabilidad, menor liquidez y mayor riesgo
- a mayor riesgo, mayor rentabilidad y menor liquidez

Teorías sobre la selección de **carteras óptimas**

Tipos de riesgos financieros según su origen

- 1) **Riesgo de liquidez:** pérdida originada por la inexistencia de una contrapartida para deshacer una posición de mercado. *Riesgo de contratación*, característico de los mercados no organizados OTC
- 2) **Riesgo Operacional:** Derivada de la existencia de anomalías o fallos en el procesamiento de información ya sean por fallos humanos, o tecnológico; también las pérdidas por información fraudulenta respecto a una operación financiera (**riesgo por fraude**) ; o por la inadecuada utilización/modelización de la información utilizada para valorar una posición de mercado (**riesgo de modelo**)
- 3) **Riesgo Legal:** surge cuando una modificación legal afecta a los términos establecidos inicialmente en una transacción (también por la inexistencia de legislación, o por laguna legal, ejemplo: deuda subordinada – *las preferentes*)

4) Riesgo de Mercado: Pérdidas derivadas del cambio adverso en el precio de un activo.

- Riesgo de **precio**
- Riesgo de **interés**
- Riesgo de **tipo de cambio**



- 1) Cómo cuantificar la probabilidad de que suban o bajen los precios de los activos financieros: Riesgo=La Varianza ¿se pueden modelizar las series financieras?
- 2) Diversificación de las Carteras para minimizar el riesgo
- 3) Estrategias de cobertura para no sufrir pérdidas en caso de variación de precios (reducir la exposición al riesgo)
- 4) Cómo cuantificar la exposición al riesgo de mercado o la máxima la pérdida esperada: Valor en Riesgo de una cartera

5) Riesgo de Crédito: Por el *incumplimiento de las obligaciones de pago* contractuales entre las partes de una operación financiera .

Scoring: puntuación, calificación o evaluación de este riesgo de crédito. Calificación crediticia del deudor (empresas, clientes, etc.). (*Riesgo Soberano*: de que sea un estado soberano el que incumpla su compromiso de pago). Calificación del riesgo de que una operación sea Fraudulenta

¿Quién realiza el scoring?.

- Departamento de Riesgos
- Agencias de calificación Standard & Poor's, Moody's (Moody's Investors Service), Fitch (Fitch Ratings, UK)
- Boureaus de créditos (préstamos al consumo, hipotecas o concesiones de tarjetas de crédito, préstamos comerciales, etc) ASNEF- Equifax, Experian-Badexcug, TransUnion, Informa (Dun&Bradstreet), RAI y otras como la central de Información de Riesgos del Banco de España

También hay modelos de calificación de fraude, de fuga de clientes o de siniestros de tráfico,...



Riesgo de Crédito y Scoring

Cuantificar el riesgo potencial de un cliente o Scoring

MODELOS DE PROBABILIDAD: MODELOS DE REGRESION CON VARIABLE DEPENDIENTE BINARIA

$$P(y=1|x)$$

- **Riesgo de crédito:** ¿Qué probabilidad existe de que si concedo un préstamo a determinado individuo no me lo devuelva?
 $P(\text{Impago}) = F(\text{características del individuo en el momento de solicitar el préstamo})$
- **Riesgo operacional:** Qué probabilidad existe de que una determinada operación bancaria sea fraudulenta
 $P(\text{Fraude}) = F(\text{características observadas de la operación y agentes que intervienen})$
- **Riesgo de fuga de clientes:** Qué probabilidad existe de que un cliente se vaya a la competencia? ¿ha qué clientes debo llamar para intentar retenerles?
 $P(\text{Fuga}) = F(\text{características observadas del cliente})$
- **Calculo de primas de un seguro** de cobertura: ¿Cuál es la prima que debo cobrarle a determinado individuo para asegurarle el caso de siniestro?
 $\text{Prima} = P(\text{siniestro}) * \text{Coste Esperado} + \text{Margen}$

$$\text{Prima} = \text{Número Esperado de Siniestros} * \text{Coste Siniestro Unitario Esperado} + \text{Margen}$$

MODELOS DE RECUENTO: Modelos Poisson, Binomial negativa, Zero truncated Poisson, Zero inflated Binomial etc $P(y=\text{num}|x)$

Credit Scoring: Valoración o Puntuación del riesgo de crédito

- **Scoring de admisión** (evaluar solicitudes para su admisión)
- **Scoring de comportamiento** (evaluar la probabilidad de incumplimiento de las operaciones ya concedidas)
- **Scoring de recobro** (evaluar la probabilidad de que una operación impagada se recupere)



Objetivo de la clase

Credit Scoring: Puntuación del riesgo de crédito

Para la modelización de las pérdidas crediticias o “Credit Losses” en realidad se necesita estimar tres elementos:

$$\text{Pérdida Esperada} = PD \times EAD \times LGD$$

Objetivo de la clase



- **PD** *Probability of Default* o **Probabilidad de incumplimiento**: probabilidad de que una contraparte no haga frente a sus obligaciones en un determinado plazo temporal.
- **EAD** *Exposure At Default* o **Exposición en incumplimiento** volumen de riesgo expuesto en el momento de incumplimiento (Cantidad total expuesta o en riesgo que se podría perder en caso de incumplimiento).
- **LGD** *Loss Given Default* o **Severidad: porcentaje final** que se pierde en caso de incumplimiento, es decir, el porcentaje no recuperado.

Nota sobre terminología:

Los **modelos de probabilidad** tienen como **variable objetivo la probabilidad de impago**. Suele definirse por tanto la variable objetivo como impago, que será una variable binaria que valdrá **1 cuando el crédito haya sido declarado como impago, y 0 en caso contrario**. Por lo tanto los modelos de probabilidad que se analizarán querrán estimar la probabilidad $P(\text{impago}=1)$. Dicha probabilidad tomará valores entre 0 y 1.

En los modelos de riesgo suele hablarse también de **buenos clientes** y **malos clientes**. Los buenos clientes serán clientes que no han hecho impago (su probabilidad de impago estimada debería ser pequeña), mientras que los malos clientes serán clientes que sí han sido declarados impagados (su probabilidad de impago debería ser alta). **Un buen modelo de crédito será aquel que sea capaz de identificar o separar bien a los buenos de los malos clientes** en función de su probabilidad estimada. Por ejemplo, un modelo que estime la misma probabilidad de impago a los buenos clientes (que han atendido correctamente todas sus obligaciones de pago) que a los malos clientes (aquellos que han incumplido con sus obligaciones de pago) será un mal modelo para medir el riesgo de crédito.

La mayoría de modelos de puntuación o valoración de riesgo de impago, sin embargo **no utilizan directamente la probabilidad de impago como medida de riesgo**. Hacen una transformación inversa de esa probabilidad de impago para convertirla en **una puntuación o nota (score)** que sirva para aprobar o suspender (denegar) los créditos. Un buen cliente (con baja probabilidad de impago) tendrá una buena puntuación o score de riesgo de crédito. Cuanto más puntuación (nota) mejor calidad del cliente. Y al contrario, un mal cliente (con alta probabilidad de impago) será aquel que tenga muy baja puntuación o score



Credit Scoring: Valoración o Puntuación del riesgo de crédito

ESTIMACION DE LA PROBABILIDAD DE IMPAGO

Se le asigna a cada individuo una puntuación o una probabilidad de que sea impagado y en función de dicha probabilidad se le concede o no el crédito.

La puntuación suele ir de menor a mayor calidad crediticia

Scorecard o Tarjeta de Puntuación del Riesgo

Variable	Atributo	Puntuación
Edad	Menor < 23	63
Edad	23-28	76
Edad	28-34	79
Edad	34-46	85
Edad	46-51	94
Edad	51- Mayor	105
Tipo Tarjeta	AMEX, VISA, Sin TRJ	80
Tipo Tarjeta	MasterCard	99
Salario	Menor <600	85
Salario	600- 1200	81
Salario	1200- 2200	93
Salario	2200 > Mayor	99
Estado Civil	Casado	85
Estado Civil	Resto	78

Ejemplo aplicación

Campo	Valor	Puntos
ID	Manuel T.	
Núm Solicitud	12345678	
Edad	43	85
Tipo TRJ	AMEX	80
Salario	1350	93
Estado Civil	Casado	85
<u>TOTAL</u>		<u>343</u>

Se establece una puntuación mínima o **Cutt-off**. Cuanto mayor puntuación mejor calidad crediticia. Hay que establecer un mínimo para aprobar

Credit Scoring: Puntuación del riesgo de crédito

Objetivo: asignar a cada individuo una puntuación o una probabilidad de que sea impagado y en función de dicha probabilidad calificarle como debajo riesgo o alto riesgo y concederle o no el crédito.

Para evaluar a cada individuo es necesario **disponer de información sobre el comportamiento de otros individuos para que por semejanza, se le pueda asignar a un nuevo cliente una probabilidad de impago** (es una medida relativa, en relación al resto de individuos).

$$P(y = 1 \mid \mathbf{x}) = P(y = 1 \mid x_1, x_2, \dots, x_k)$$

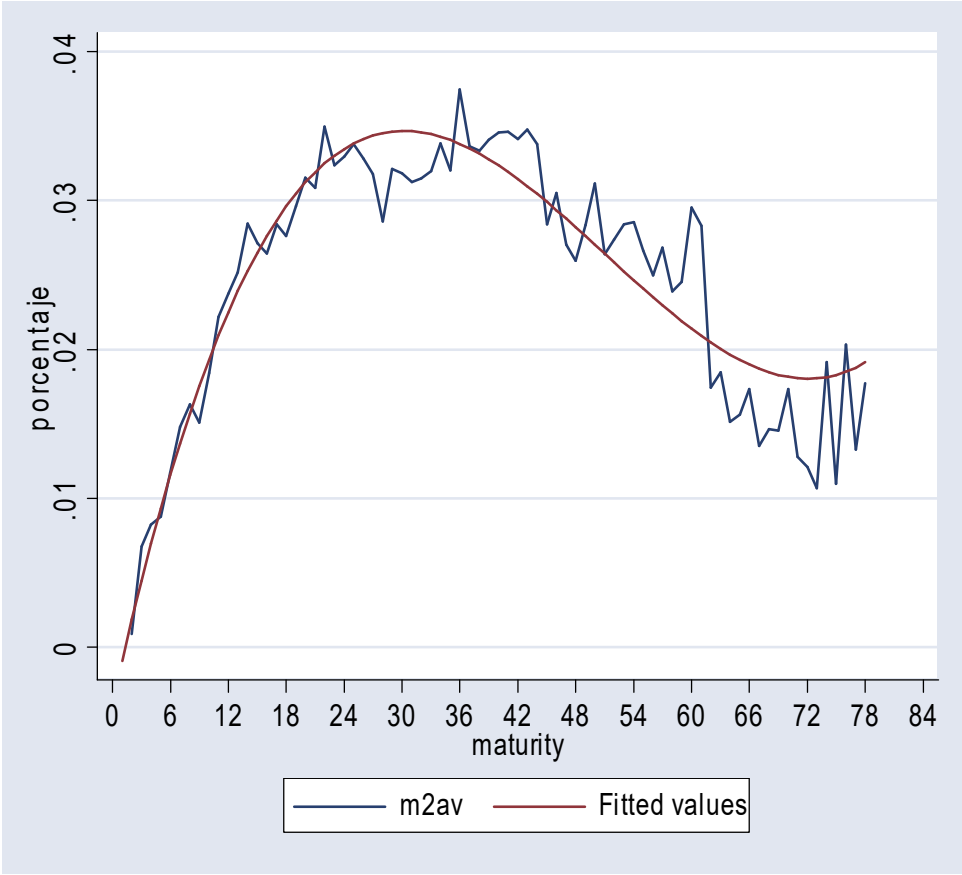
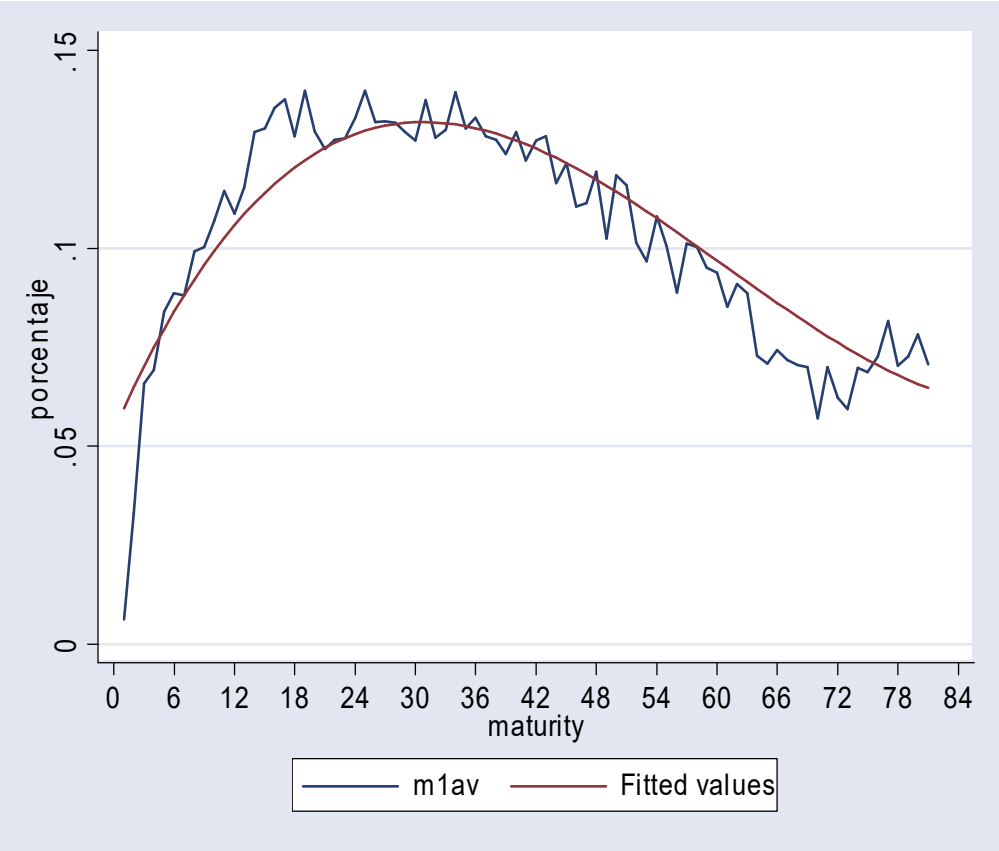
Hay que **entrenar** al modelo con individuos para los que se haya observado que hayan **impagado** (clientes fugados, operaciones fraudulentas, siniestros observados...). Por eso es muy importante analizar los **sesgos de selección muestral**, esto es que la muestra de individuos que estés utilizando para entrenar el modelo sea similar y representativa de la población sobre la que después aplicarás tu modelo

Fases del análisis

- 1) Selección de datos de clientes, de sus características y de si resultaron impagados o no
 - **Depuración de datos:** (Estadística univariante) ¿Datos Faltantes Missing? ¿Hay que imputar? ¿se dejan como otra categoría? ¿Datos extremos anómalos? ¿transformación de variables continuas (ej: logaritmos)?
 - **Definición de la Variable Objetivo : Impago** ¿Qué ventana temporal se utiliza para observar impagos? ¿Cuánto tiempo tiene que pasar para ser considerado impagado?
 - **Selección y transformación de variables explicativas** ¿Qué variables seleccionamos? , ¿transformamos variables continuas?. ¿agrupación de variables? Se pueden utilizar criterios de información , WOE, KS, estadísticos bi-variantes , o modelos de estimación automáticos por pasos. Ojo con la Sobreparametrización o con el sobreajuste

- 2) Estimación de modelos de regresión de probabilidad multivariante y construcción de las tarjetas de puntuación (**Regresión logística**) (interpretación de los resultados)
- 3) ¿**Sesgos**? **Omisión de variables relevantes** relacionadas con otras variables explicativas. **Errores de medida**. **Inferencia de denegados (sesgo de autoselección)**
- 4) **Validación del Modelo**. La muestra suele dividirse en dos, una parte será la que se utilice para el entrenamiento del modelo (estimar el modelo de regresión), y la otra para validar el modelo (ver si las previsiones de impago son apropiadas). **Validación cruzada** para hacer diferentes divisiones
- 5) Obtención del **modelo final**, construcción de la tarjeta de puntuación y establecimiento del punto de corte o nota de referencia para aprobar los créditos

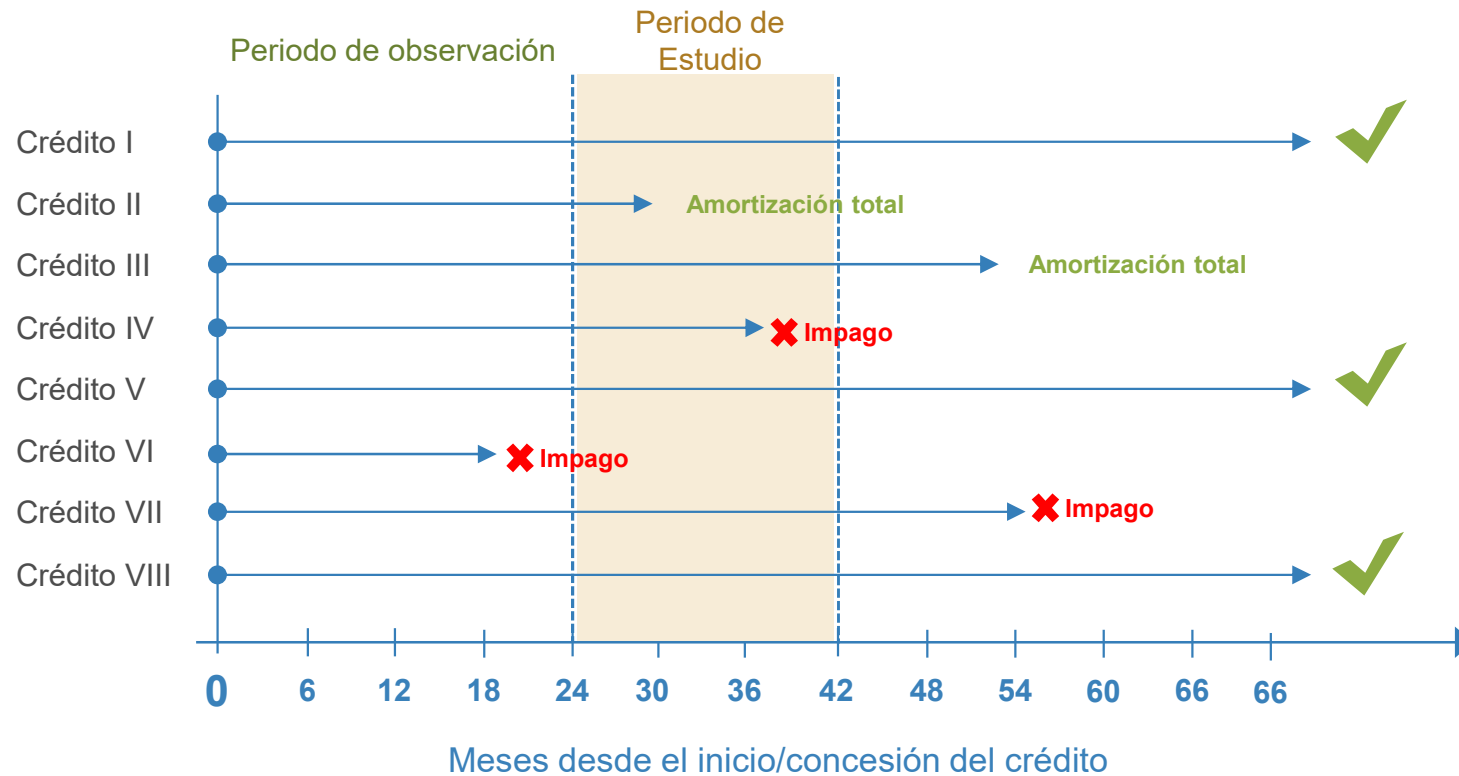
DETERMINACION DE LA VARIABLE OBJETIVO

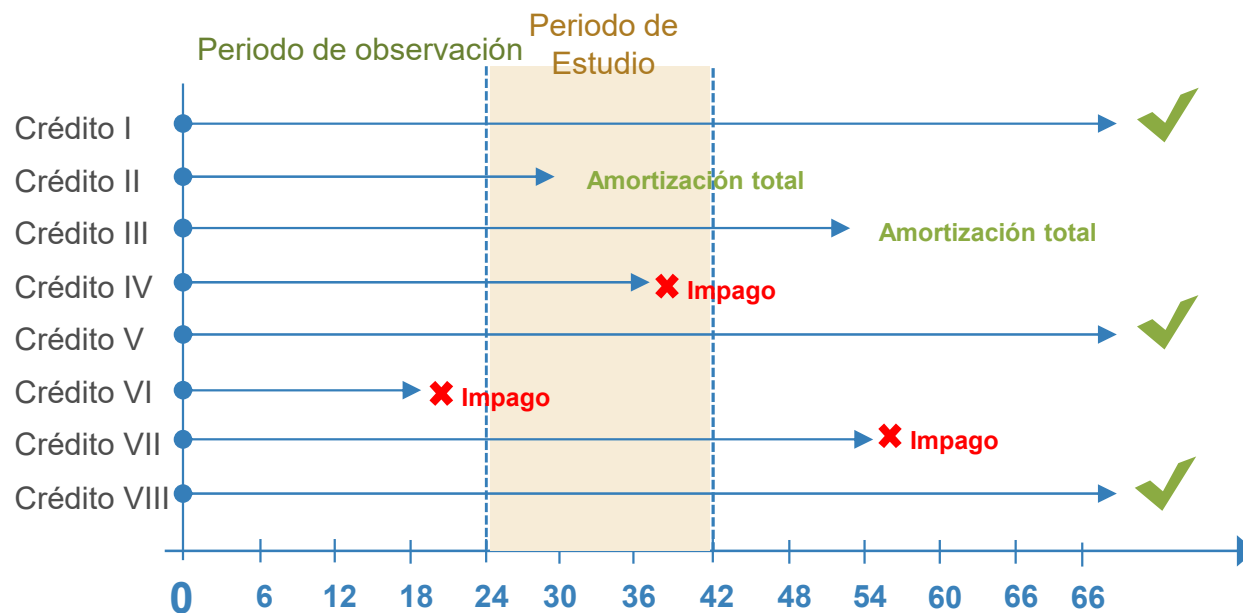


M1	Mora >= 1 día ; 1 cuota representa entre 1 y 30 días de retraso
M2	Mora >= 31 días ; 2 cuotas representan entre 31 y 60 días de retraso
M3	Mora >= 61 días ; 3 cuotas representan entre 61 y 90 días de retraso
M4	Mora >= 91 días ; 4 cuotas representan entre 91 y 120 días de retraso
M5	Mora >= 121 días ; 5 cuotas representan entre 121 y 150 días de retraso
M6	Mora >= 151 días ; 6 cuotas representan mayor a 151 días de retraso

variable	Obs	Mean	Std. Dev.
m1av	106	.0949663	.0291736
m2av	105	.0253732	.009985
m3av	104	.0102153	.0047353
m4av	103	.0065403	.0032429
m5av	102	.0055776	.0030263

Creación de la tabla de datos inicial de **clientes buenos** y **clientes malos**





Créditos descartados para el estudio:

- Crédito II
- Crédito VI

Créditos declarados como malos para el estudio (impagados):

- Crédito IV

Créditos declarados como buenos para el estudio (en la ventana de observación):

- Crédito I
- Crédito III
- Crédito V
- Crédito VII
- Crédito VIII

Obs/Var	X_1	X_2	...	X_k	Impago
Crédito I					0
Crédito III					0
Crédito IV					1
Crédito V					0
Crédito VII					0
Crédito VIII					0

Aunque el Crédito VII finalmente resultó impagado, lo hizo después de la ventana de observación

Selección y transformación de variables Explicativas

Existen diferentes métodos que pueden aplicarse para hacer **una selección inicial** de posibles variables explicativas o independientes a introducir en el modelo.

Descarte inicial de variables:

- Variables con **excesivo** número de datos perdidos (NA)
- Variables con **excesivos valores idénticos**
- Demasiados **Datos extremos** (¿son atípicos?)

Depuración: Análisis de Datos Perdidos y Valores Extremos:

- ¿Qué hacemos con los **datos atípicos**? ¿los **convertimos en NA**?
- ¿Qué hacemos con los **datos perdidos** (NA)?
 - **Imputamos** valores perdidos?
 - ¿Convertimos los valores perdidos en una **nueva categoría** dentro de la variable?

Se **seleccionarán variables que a priori muestren cierta asociación con la variable objetivo**. Para contrastar la existencia de asociación entre la variable objetivo (impago) pueden utilizarse

a. **Estadísticos tradicionales de asociación bivalente** (entre la variable objetivo y cada una de las potenciales variables explicativas), por ejemplo:

- **Chi cuadrado** (cuando la **variable explicativa también es categórica**)
- **Test ANOVA** de diferencia de medias (cuando la **variable explicativa sea numérica**)

Como hay muchos datos, todo sale significativo

b. Estadísticos de asociación basados en **criterios de concentración**

cuantifican cómo están concentrados los buenos y los malos clientes, los ceros y los unos de mi variable objetivo, en cada una de las categorías de la variables explicativas:

- Valor de la información: Media ponderada del Woe
- Gini

Selección de variables Explicativas según criterios de concentración

Estos estadísticos buscan variables que contengan categorías que concentren relativamente a muchos malos clientes ($impago=1$) o a muchos buenos clientes ($impago=0$), de forma que dichas categorías proporcionen información relevante para discriminar o separar a los buenos de los malos clientes.

Ejemplo: una variable cuyas categorías presente el mismo porcentaje de buenos y malos clientes en todas sus categorías, en realidad no aporta información para separar a los buenos de los malos. En caso de que un nuevo cliente caiga en alguna de esas categorías no sabremos si asignarlo al grupo de buenos o al de malos clientes. Por el contrario, una variable que tenga alguna categoría que concentre a muchos malos o a muchos buenos clientes, si servirá para predecir si un nuevo cliente que cae justo en esa categoría será un mal o buen cliente (respectivamente)

En la literatura se han establecido dos estadísticos que utilizan este concepto de concentración que pueden utilizarse para hacer una selección previa de variables a incluir en el modelo:

- Valor de información (*Information Value* o *IV*)
- Índice de Gini.

Además se han establecido una serie de criterios o reglas prácticas para determinar cuando incluir o no una variable en el modelo en función del valor que tomen estas medidas de concentración que no se basan en contrastes tradicionales, ni utilizan el concepto de *p-valores* de la inferencia estadística tradicional, por lo que son especialmente recomendables para el caso de modelos en los que haya que trabajar con grandes volúmenes de datos

Valor de información (Information Value)

El **valor de información** o **IV** por sus siglas en inglés (*Information Value*) es un estadístico que proporciona información sobre la concentración relativa de malos y buenos clientes en las diferentes categorías de una variable categórica. Cuanto mayor sea el IV de una variable explicativa, mayor concentración en sus categorías y por tanto mayor información proporciona esa variable para separar a los buenos de los malos clientes.

El IV se calcula como una media ponderada de otra medida de concentración denominada **Peso de la Evidencia** o **WoE** por sus siglas en inglés (*Weight of Evidence*) que es un estadístico que se calcula para cada una de las categorías que conforman una variable categórica

WoE (weight of evidence): Es una medida de la fuerza de las categorías de una variable para separar a los buenos clientes (G del inglés Good) de los malos clientes (B del inglés Bad). Definida como el logaritmo neperiano del porcentaje de malos frente al porcentaje de buenos clientes que concentra una categoría

$$WoE_i = \ln(B_i/B) - \ln(G_i/G) \quad \text{para cada categoría } i$$

Ejemplo: **WoE (weight of evidence):** $WoE_i = \ln(B_i/B) - \ln(G_i/G)$ para cada categoría i

EDAD	TOTAL	%	Buenos	%BUENOS	Malos	%MALOS	WOE	IV
Perdidos	1000	2.5%	860	2.4%	140	3.6%	0.43	0.005
18-22	4000	10.0%	3040	8.4%	960	25.0%	1.09	0.181
23-26	6000	15.0%	4920	13.6%	1080	28.1%	0.73	0.105
27-29	9000	22.5%	8100	22.4%	900	23.4%	0.05	0.000
30-35	10000	25.0%	9500	26.3%	500	13.0%	-0.70	0.093
36-44	7000	17.5%	6800	18.8%	200	5.2%	-1.28	0.175
más de 44	3000	7.5%	2940	8.1%	60	1.6%	-1.65	0.108
TOTAL	40000	100.0%	36160	100.0%	3840	100.0%	0.00	0.668

- El Tramo de edad de **18 a 22** concentra al **8.4% del total de buenos** pero al **25% del total de malos**. Concentra relativamente a un alto porcentaje de malos. Tiene un WOE muy alto de 1.09, indicando que la cantidad de malos clientes (B) es muy alta en relación a la de buenos clientes (G)
- El tramo central de **27 a 29 años** concentra aproximadamente al mismo porcentaje de buenos (22.4%) que de malos (23.4%), por lo que esta categoría proporciona poca evidencia de concentración de buenos o malos, no servirá para separar a los buenos de los malos (WoE= 0.05)
- Por último, los tramos de mayor edad concentran a muchos buenos clientes en relación a los malos, por lo que tendrán un WoE negativo tanto más elevado cuanto mayor la presencia de buenos frente a malos en dicha categoría.



Para la selección de variables potencialmente explicativas para estimar el modelo, lo que se busca son variables que concentren en sus categorías a muchos buenos o a muchos malos en términos relativos, porque ayudarán a separar a los buenos de los malos. Un indicador global de esa concentración en las diferentes categorías de una variable es el Valor de la Información de la variable:

IV (Information Value): Es una medida del poder global de una variable para discriminar entre Buenos y Malos y se construye como una media ponderada de los WoE de cada categoría

$$IV = \sum (B_i/B - G_i/G) * WoE_i$$

siendo i cada una de las categorías que conforman la variable

Nótese que las ponderaciones vienen dadas por la diferencia simple entre el porcentaje de malos(B) y buenos (G), que tendrá el mismo signo que su WoE, por lo que cuanto mayor concentración de buenos o malos clientes existan en las categorías de una variable mayor será el valor del IV de ésta

EDAD	TOTAL	%	Buenos	%BUENOS	Malos	%MALOS	WOE	IV
Perdidos	1000	2.5%	860	2.4%	140	3.6%	0.43	0.005
18-22	4000	10.0%	3040	8.4%	960	25.0%	1.09	0.181
23-26	6000	15.0%	4920	13.6%	1080	28.1%	0.73	0.105
27-29	9000	22.5%	8100	22.4%	900	23.4%	0.05	0.000
30-35	10000	25.0%	9500	26.3%	500	13.0%	-0.70	0.093
36-44	7000	17.5%	6800	18.8%	200	5.2%	-1.28	0.175
más de 44	3000	7.5%	2940	8.1%	60	1.6%	-1.65	0.108
TOTAL	40000	100.0%	36160	100.0%	3840	100.0%	0.00	0.668

Criterio para la selección de variables (Siddiqi, 2006)

Si IV es menor que 0.02, la variable es no predictiva, si es mayor que 0.3 es muy predictiva

Regla de aplicación

- Si es menor que 0,02 -> Variable no predictiva
- Si está entre 0,02 y 0,1 -> Influencia débil
- Si está entre 0,1 y 0,3 -> Influencia media
- 0,3 y mayor -> Fuerte influencia.



Índice de GINI

Otra medida de concentración de buenos o malos clientes que también se usa en la literatura para la selección de variables es el índice de Gini, que toma valor 0 si una variable tiene el mismo porcentaje de malos (**B**) en relación al porcentaje de buenos (**G**) en todas sus categorías; y que valdrá 1 si todos los malos se concentran en una sola categoría de la variable (concentración máxima)

Para construir el índice de Gini primero deben ordenarse las m categorías en orden descendente por la proporción de malos en cada uno de ellas. Esto es importante porque la fórmula del índice de Gini requiere calcular frecuencias acumuladas.

$$Gini = \left(1 - \frac{2 \times \sum_{i=2}^m \left(B_i \times \sum_{j=1}^{i-1} G_j \right) + \sum_{i=1}^m (B_i \times G_i)}{B \times G} \right) \times 100$$

Criterio para la selección de variables (Siddiqi, 2006): que exista un mínimo de concentración de malos, un índice de Gini superior a 0.15 Sas: Mínimo 0.20

No es una regla pero se recomienda:

- Entre 0-5% No usar la variable en el modelo multivariante
- Entre 5%-15% Poder bajo pero se puede usar la variable en el modelo multivariante
- >15% Poder razonable se recomienda usar la variable en el modelo multivariante

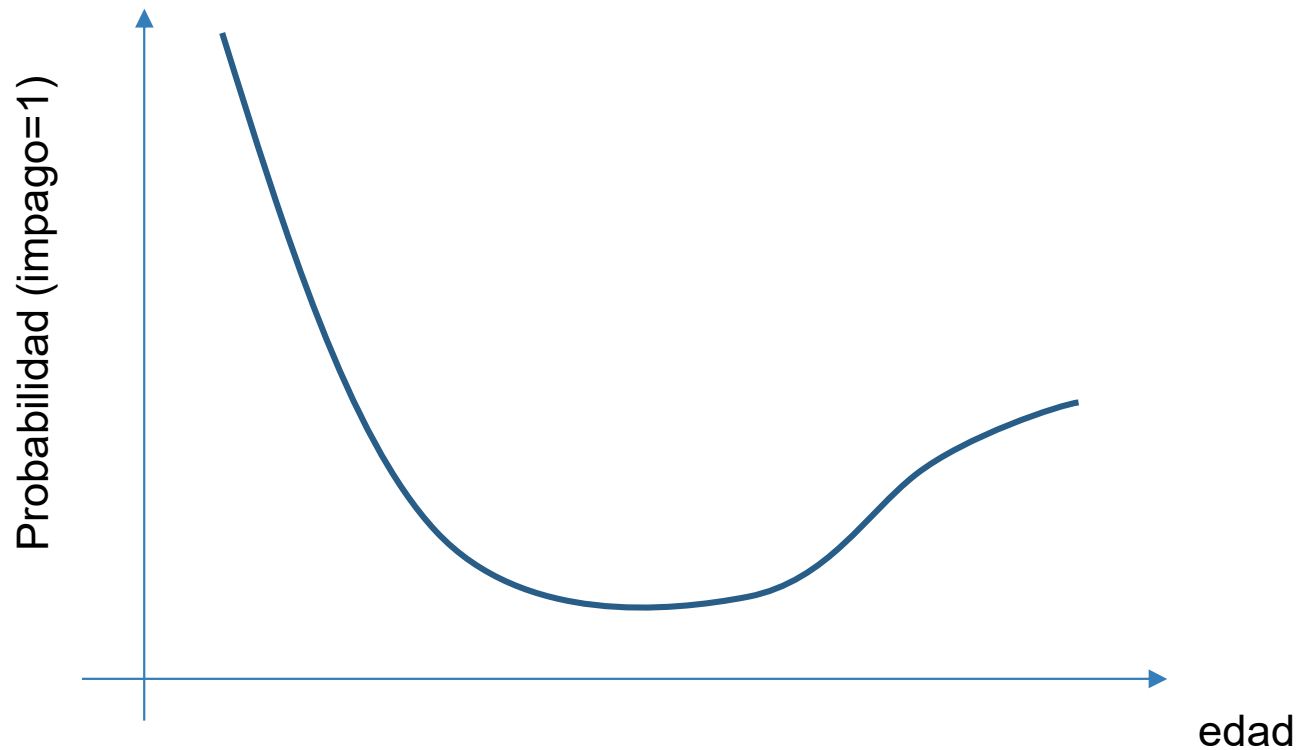
Categorización de Variables Continuas (binning) y agrupación de categorías

La aplicación tanto del **IV** como del índice de **Gini** requiere de la existencia de diferentes categorías dentro de una variable, esto es, requiere que todas las potenciales variables explicativas sean categóricas, por lo que se requiere **reconvertir a categóricas todas las variables continuas**. Este proceso se denomina categorización o *tramificación o binning* de las variables categóricas.

Frente a la pérdida de información que puede suponer tramificar o categorizar una variable continua, su principal ventaja (además de poder utilizar el IV o Gini como criterio de selección de variables) es la de permitir **captar las no linealidades** que puedan existir entre las variables continuas y la variable **objetivo** (sobre todo cuando se utilizan modelos lineales de probabilidad)

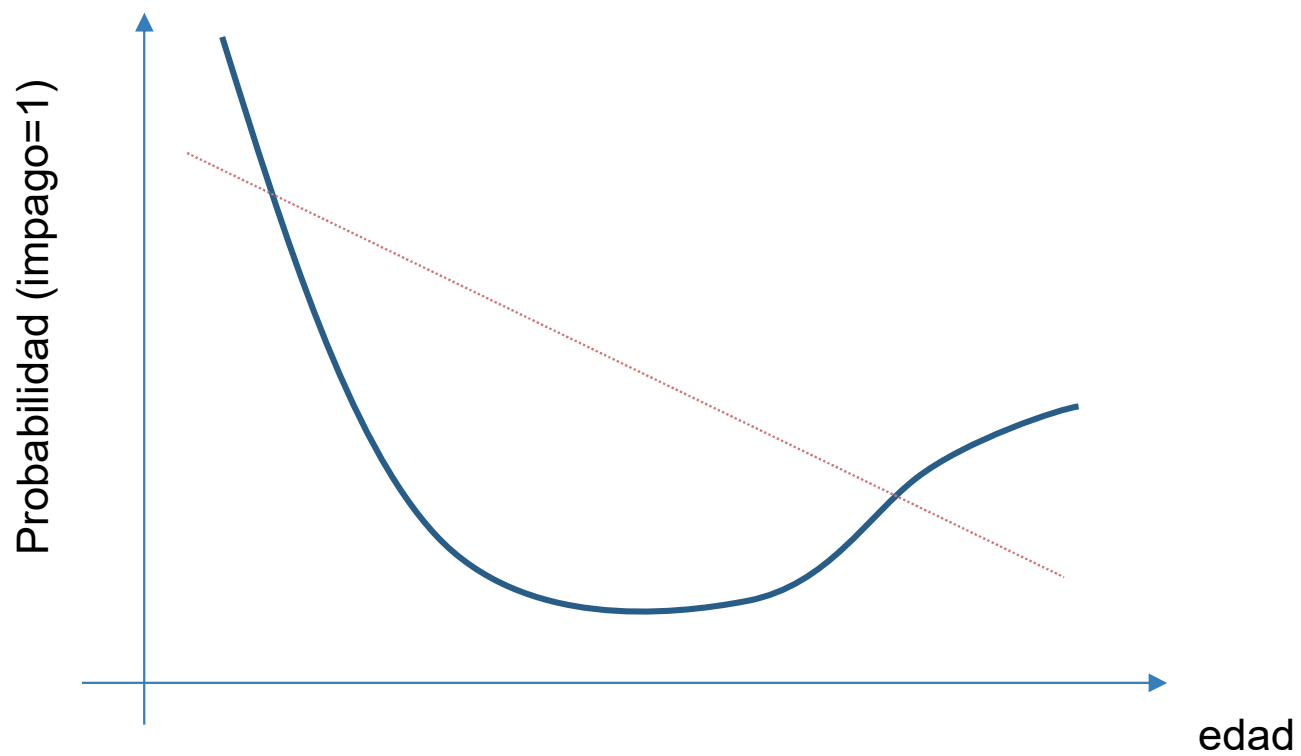
Ejemplo relación de la edad con la probabilidad de impago

Suponga que la **relación entre la edad y la probabilidad de impago es no lineal**: los más jóvenes son más propensos al impago; a medida de que se van cumpliendo años nos volvemos mejor pagadores y se reduce la probabilidad de impago; y ya a más avanzada edad vuelve a subir ligeramente la probabilidad de impago



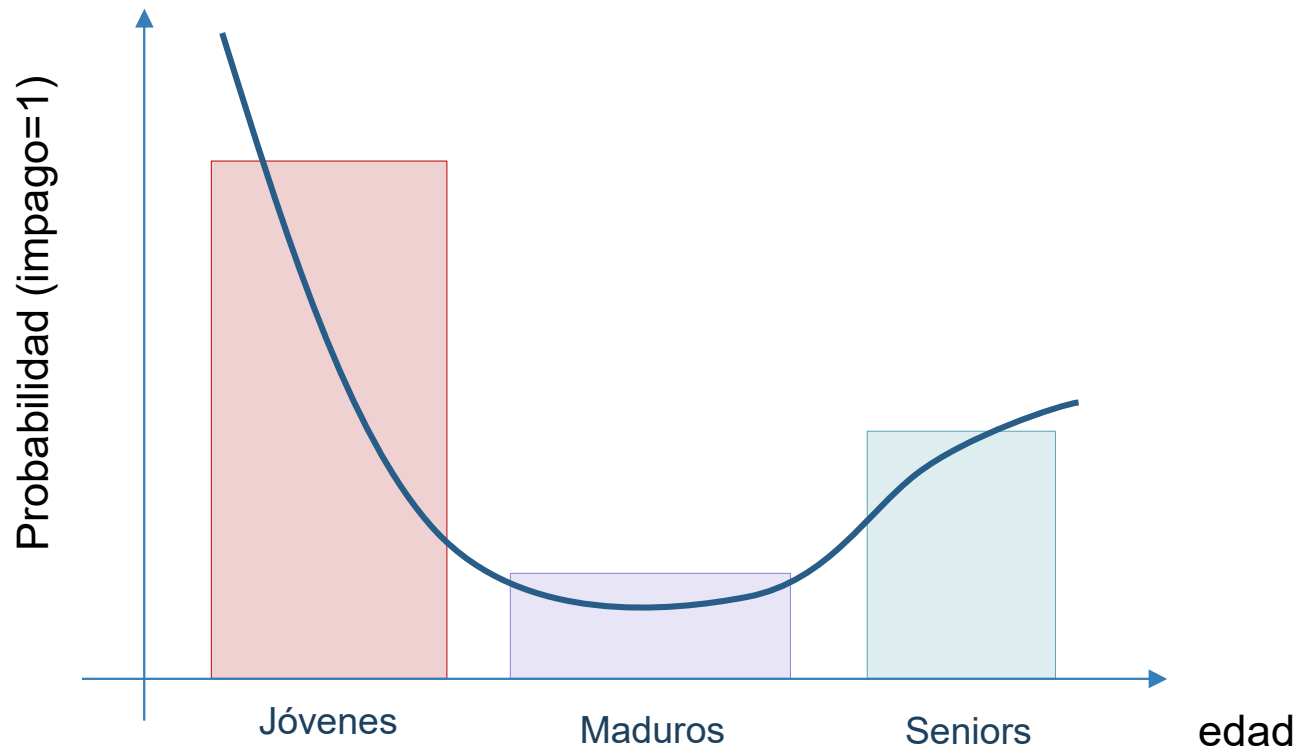
Ejemplo relación de la edad con la probabilidad de impago

No podemos ajustar esta relación entre **edad** y la **probabilidad de impago** mediante un **modelo lineal**.



Ejemplo relación de la edad con la probabilidad de impago

Así que una forma de captar la no linealidad mediante un modelo lineal es convirtiendo la variable continua edad en una variable categórica de forma que puedan utilizarse estas categorías como variables binarias diferentes (una para cada tramo de edad) para captar así la relación no-lineal original entre probabilidad de impago y edad



Agrupación de categorías

La conversión de variables continuas en variables categóricas o tramificadas permite (y es condición necesaria) aplicar las medidas de concentración IV y GINI para seleccionar variables que potencialmente van a ayudar a separar a los buenos de los malos clientes, y por tanto a variables potencialmente buenas candidatas a ser incorporarlas al modelo de probabilidad como variables explicativas.

Existen diferentes técnicas para hacer esta **tramificación de variables continuas**. Pero una de ellas es precisamente la de **elegir el número de tramos o categorías que maximiza el valor de información** (o índice de GINI).

Puede procederse, por ejemplo, de la siguiente forma. Se ordena de menor a mayor la muestra en función de la variable numérica que se quiere categorizar. Se hacen 20 grupos, categorías o tramos iniciales con el mismo número de registros en cada grupo y se calcula para cada categoría su WoE y el IV total de la variable. A continuación se van reagrupando categorías con WoE similar de forma que se vaya reduciendo el número total de grupos o categorías a la vez que se consigue aumentar el IV final de la variable

Este mismo criterio de ir agrupando categorías que proporcionen la misma información de forma que el IV total de la variable vaya aumentando puede utilizarse también para **recodificar o reagrupar categorías en variables categóricas** (reducir el número de categorías que tendrán las variables categóricas)

La transformación WOE de las Variables categóricas

El proceso de construcción de los modelos de probabilidad, ha requerido la tramificación o categorización de variables continuas. Con todas las variables categorizadas se ha procedido a analizar los índices de concentración y en su caso reagrupación de tramos de forma que la cantidad de información que proporcionen dichas categorías sea la máxima para poder discriminar entre buenos y malos clientes.

Una vez seleccionadas las variables que se introducirán a priori en el modelo de probabilidad, en la metodología de riesgos suele realizarse una transformación adicional a las variables seleccionadas: **convertir las variables categóricas en variables continuas WOE**

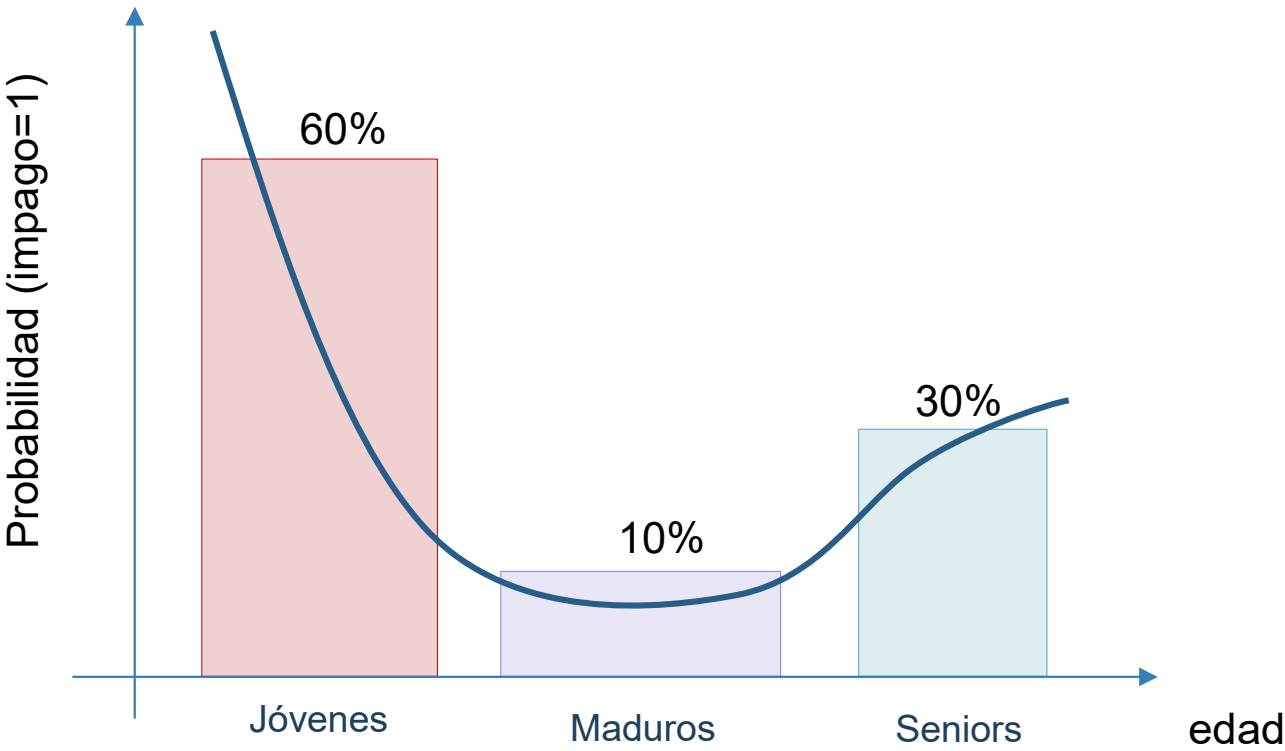
¿Para qué? para no tener que construir una variables dicotómica para cada una de las categorías de una variable y reducir así el tamaño de la matriz de variables explicativas y ahorrar tiempo de computación en el proceso de estimación de los modelos o algoritmos de estimación y validación de los modelos

Además, el uso de los **WOE permite linealizar todas la relaciones no lineales**, de forma que se mejore la potencia de modelos lineales. Cuando existen no linealidades entre la variable original y la variable objetivo probabilidad de impago, dichas no linealidades se convierten en lineales con la **transformación WOE**. De esta forma es posible tratar las relaciones como esencialmente lineales (usando un modelo lineal “generalizado” de probabilidad, sin pérdida de información, que no sería válido si se utilizase la variable continua original)

Ejemplo relación de la edad con la probabilidad de impago

Volviendo al anterior ejemplo supongamos los siguientes datos de frecuencia de impagados en cada tramo de edad

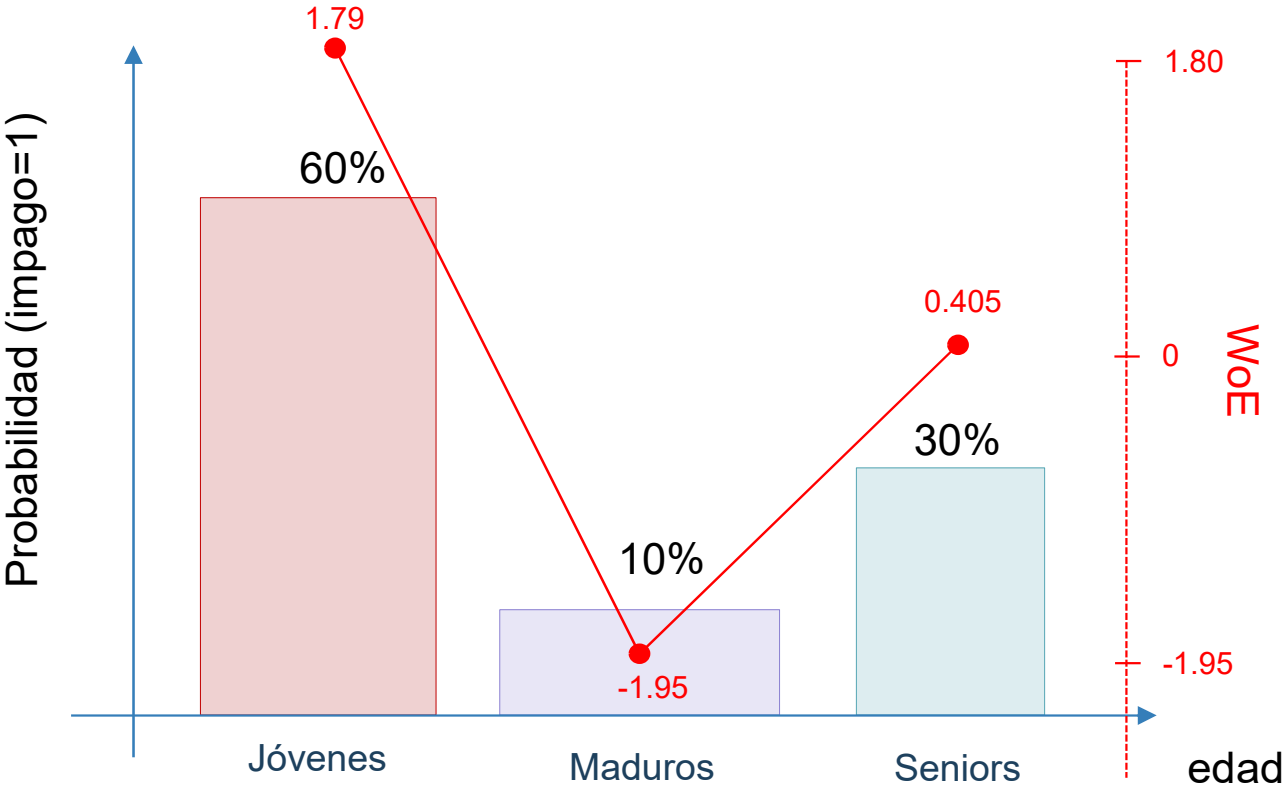
	Porcentaje de Malos	Porcentaje de buenos	WOE
Jóvenes	60%	10%	1.792
Maduro	10%	70%	-1.946
Seniors	30%	20%	0.405



Ejemplo relación de la edad con la probabilidad de impago

Represetamos los WoE de cada categoría en rojo y en escala de la derecha

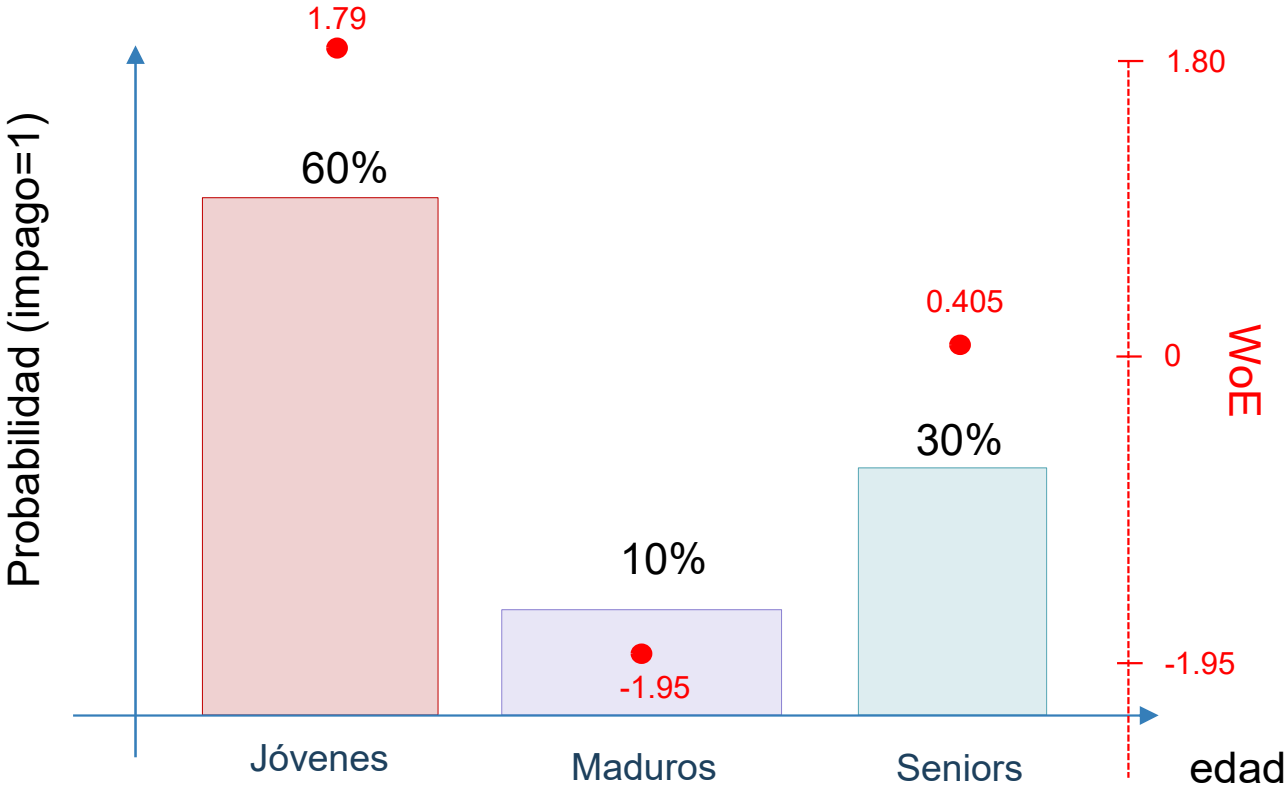
	Porcentaje de Malos	Porcentaje de buenos	WOE
Jóvenes	60%	10%	1.792
Maduro	10%	70%	-1.946
Seniors	30%	20%	0.405



Ejemplo relación de la edad con la probabilidad de impago

Procedemos a asignar a cada categoría su valor WoE y dibujamos no respecto a edad sino a WoE en el eje x

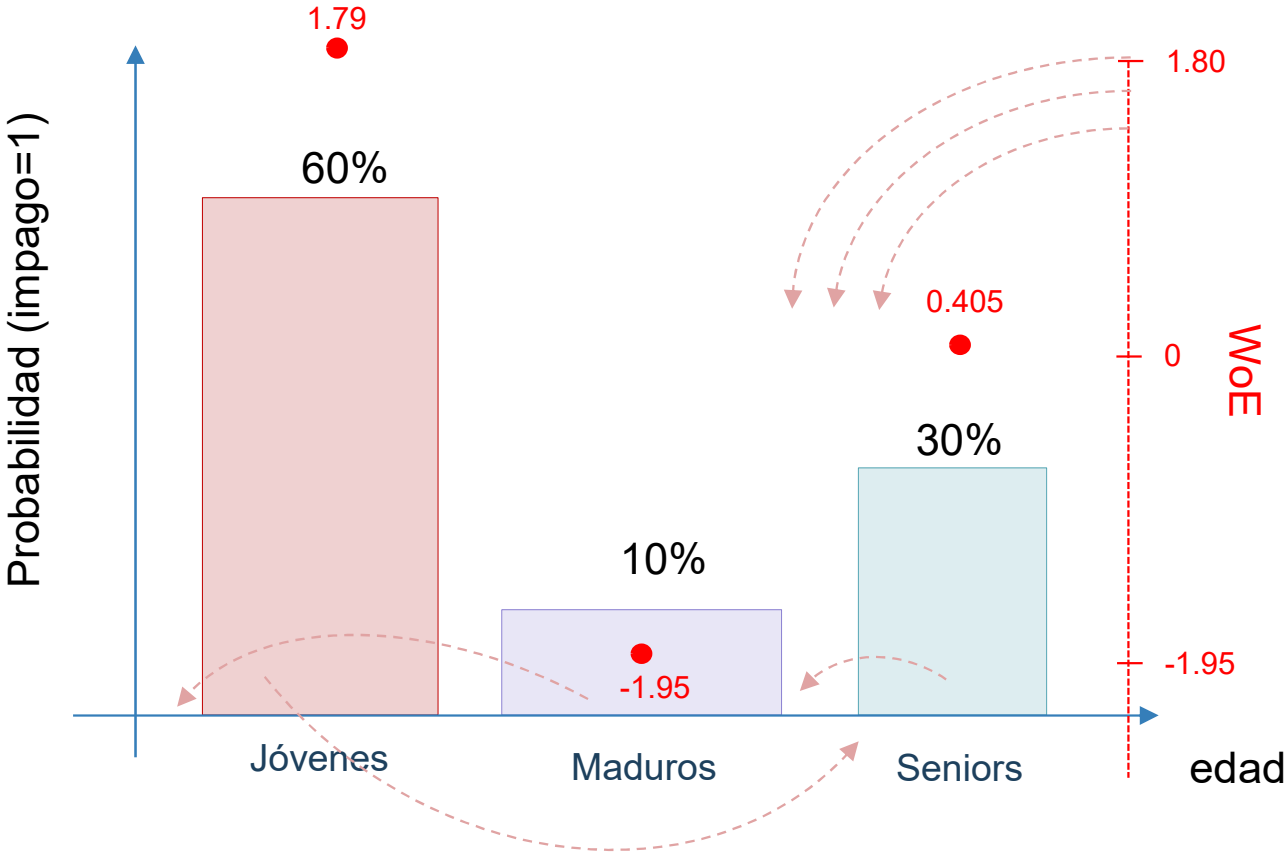
	Porcentaje de Malos	Porcentaje de buenos	WOE
Jóvenes	60%	10%	1.792
Maduro	10%	70%	-1.946
Seniors	30%	20%	0.405



Ejemplo relación de la edad con la probabilidad de impago

Procedemos a asignar a cada categoría su valor WoE y dibujamos no respecto a edad sino a WoE en el eje x

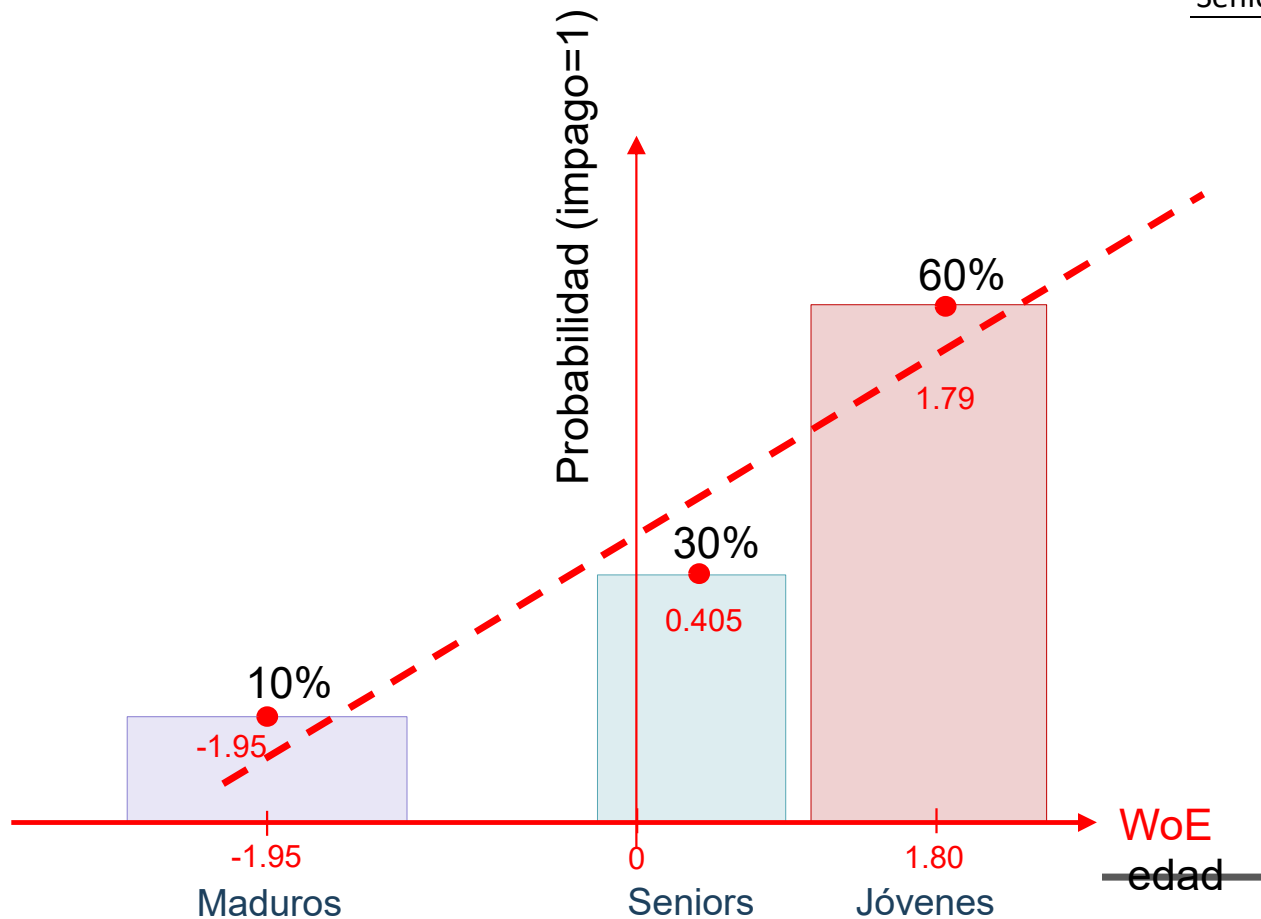
	Porcentaje de Malos	Porcentaje de buenos	WOE
Jóvenes	60%	10%	1.792
Maduro	10%	70%	-1.946
Seniors	30%	20%	0.405



Edad	Edad Woe
Jóven	1.792
Jóven	1.792
Senior	0.405
Maduro	-1.946
Senior	0.405
...	...
Maduro	-1.946
Senior	0.405

Ejemplo relación de la edad con la probabilidad de impago

Procedemos a asignar a cada categoría su valor WoE y dibujamos no respecto a edad sino a WoE en el eje x



	Porcentaje de Malos	Porcentaje de buenos	WOE
Jóvenes	60%	10%	1.792
Maduro	10%	70%	-1.946
Seniors	30%	20%	0.405

Edad	Edad Woe
Jóven	1.792
Jóven	1.792
Senior	0.405
Maduro	-1.946
Senior	0.405
...	...
Maduro	-1.946
Senior	0.405

RESUMEN:

El procedimiento para la **selección de variables utilizando medidas de concentración** (Valor de información – **IV** - o índice de **GINI**) **requiere:**

1. Categorización de variables continuas (**binning**)
2. Agrupación de categorías (**se busca que cada variable tenga el menor número de categorías posibles de forma que éstas proporcionen la mayor cantidad de información**)
3. Selección de variables según criterio **IV>0.02** o **GINI > 0.15**
4. Construcción de Variables WOE (**se vuelven a recodificar las variables categóricas para que puedan tratarse como si fueran variables continuas**)

A partir de aquí se puede comenzar a estimar el modelo de probabilidad:

5. **Estimación del modelo de regresión logística (con las variables WoE como explicativas)**
6. **Validación y diagnosis (análisis de la posible existencia de sesgos de selección muestral)**
7. **Obtención del modelo final**
8. **Transformar los odds ratios en una puntuación o score y establecer punto de corte y la tarjeta de puntuación**

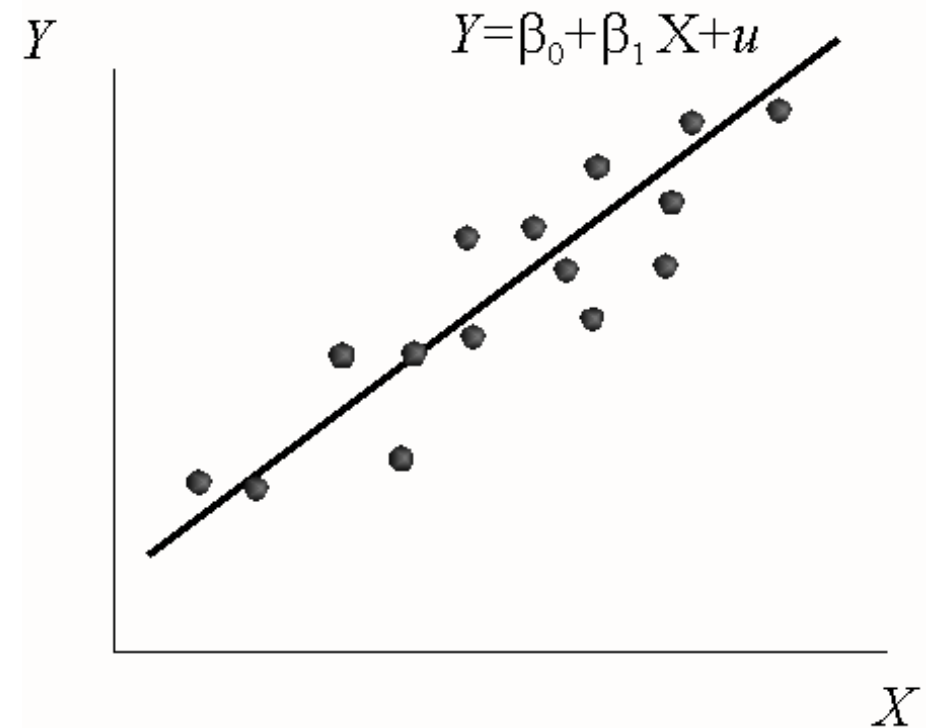
Ejemplos de variables que frecuentemente se encuentran en modelos de crédito al consumo.

- ✓ Edad
- ✓ Sexo
- ✓ Educación
- ✓ Estado Civil
- ✓ Estatus residencial
- ✓ Tiempo viviendo en el domicilio
- ✓ Actividad
- ✓ Puesto en el trabajo
- ✓ Tiempo trabajando
- ✓ Tiempo como Cliente en el Banco
- ✓ Industria en la que trabaja

Una vez que tenemos Seleccionadas las Variables: estimación de modelos de “clasificación” o de regresión de probabilidad multivariante, diagnosis e interpretación de los resultados y construcción de las tarjetas de puntuación

MODELOS de Clasificación (o modelos de Probabilidad):

- Modelo Lineal de probabilidad
- **Regresión logística**
- modelo Probit
- árboles de decisión
- Random Forest
- Gradieng/Adaptative Boosting
- Suport Vector Machines
- Redes Neuronales
- Otros (**Data ROBOT**)



y representa la variable impago $\left\{ \begin{array}{l} \text{Si } y=1 \text{ el cliente ha incurrido en impago} \\ \text{Si } y=0 \text{ el cliente no ha incurrido en impago (buen cliente)} \end{array} \right.$

Para predecir si un determinado cliente va a incurrir en pago si le concedo un préstamo, o si por el contrario no incurrirá en impago, estimamos un modelo de probabilidad, donde la **variable dependiente** es, en realidad, **la probabilidad de incurrir en impago**

$$P(y = 1 \mid \mathbf{x}) = P(y = 1 \mid x_1, x_2, \dots, x_k)$$

Propiedades

variable binaria $y=\{0,1\}$ entonces

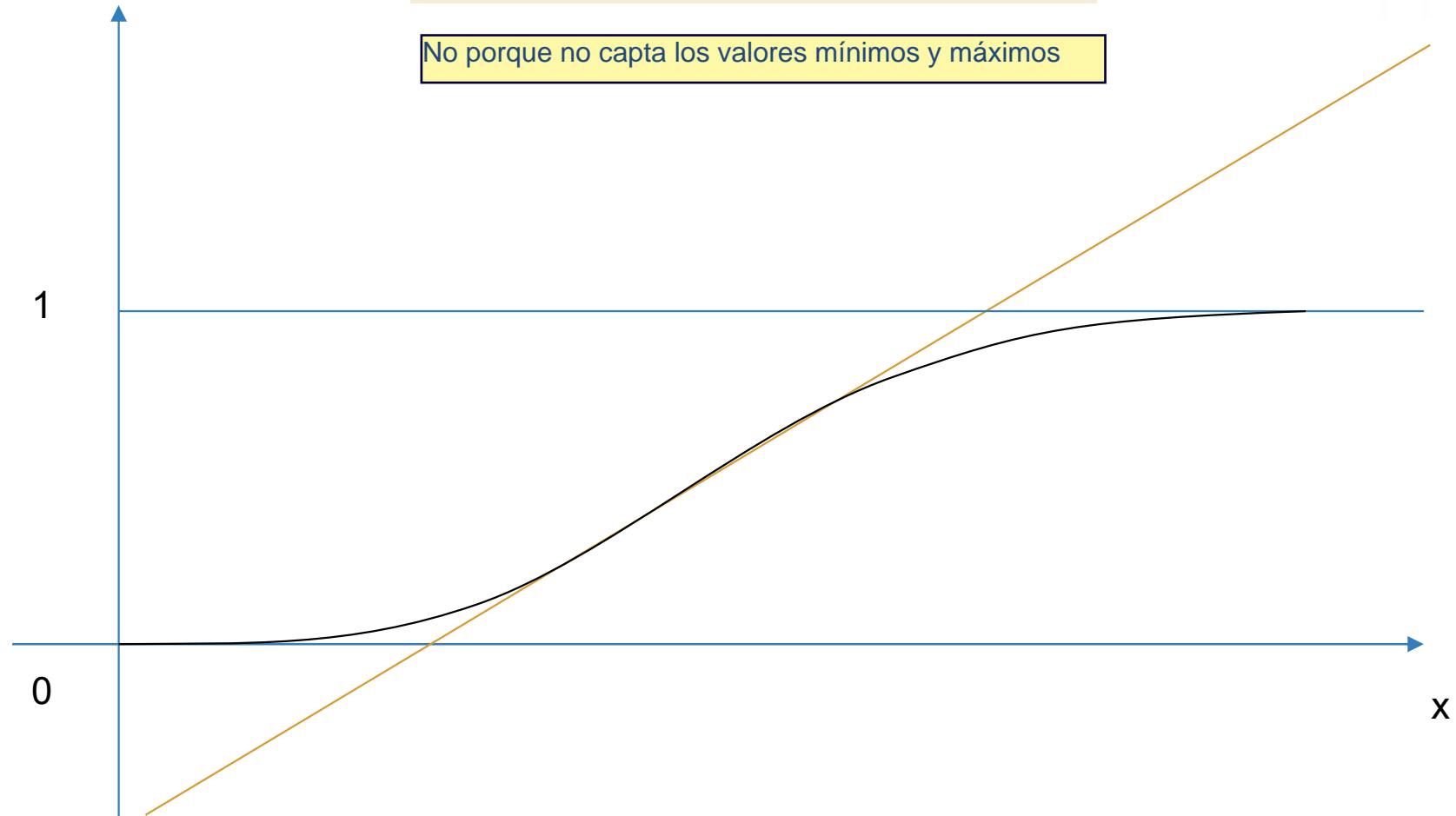
- $P(y=1)=P$,
- $P(y=0)=1-P$,
- $E(y)=1 \cdot P(y=1)+0 \cdot P(y=0)=P$
- $Var(y)=P(1-P)^2+(1-P)(0-P)^2=P(1-P)^2+(1-P)(P)^2=P(1-P)[(1-P)+P]=P(1-P)$

$$P(y = 1 \mid \mathbf{x}) = P(y = 1 \mid x_1, x_2, \dots, x_k)$$

¿Se puede utilizar una especificación lineal? (modelo lineal de probabilidad)

No porque no capta los valores mínimos y máximos

Probabilidad ($y=1 \mid x$)




Modelos de Regresión con Variable dependiente Binaria o modelos paramétricos de Probabilidad

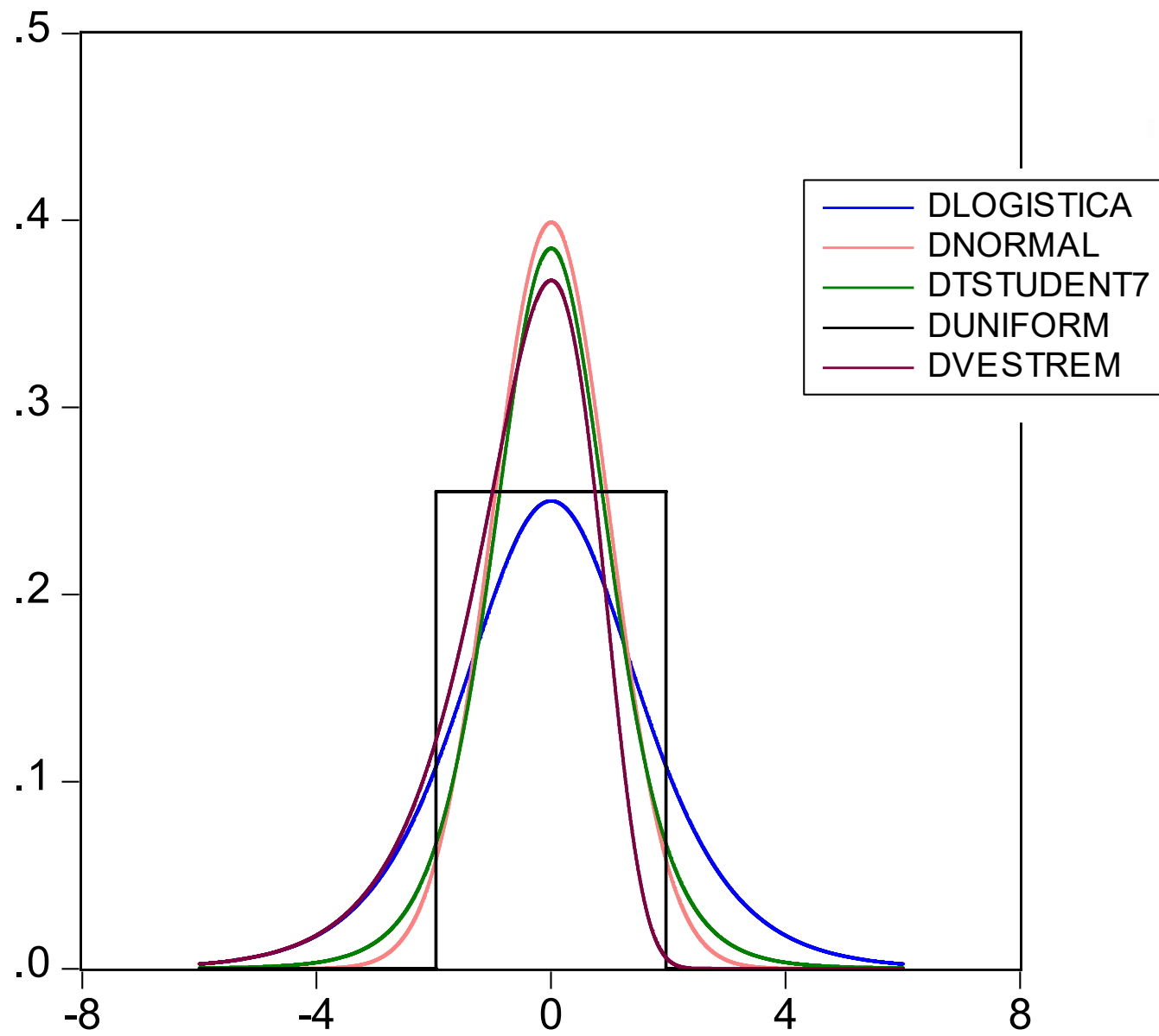
$$P(y = 1 \mid \mathbf{x}) = P(y = 1 \mid x_1, x_2, \dots, x_k)$$

Son **modelos lineales generalizados** (GLM de sus siglas en inglés)

$$P(y = 1 \mid \mathbf{x}) = P(y = 1 \mid x_1, x_2, \dots, x_k) = G(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$$

¿G() Función de Distribución
de Probabilidad?

- 
- Modelo Lineal de Probabilidad: ***Uniforme***
 - Modelo Probit: ***Normal Estándar***
 - Modelo Logit: ***Logística***
 - Modelo Gompit: ***Valor Extremo***



$$P(y = 1 | \mathbf{x}) = P(y = 1 | x_1, x_2, \dots, x_k) = G(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$$

Los modelos de elección binaria se derivan del denominado modelo de variable latente subyacente que satisface las suposiciones del modelo lineal clásico (Wooldrige, 2001, pág.532). Sea y^* una variable no-observable o variable latente determinada por

$$y^* = \beta_0 + \mathbf{x}\boldsymbol{\beta} + e \quad (3)$$

Este modelo (3) no puede estimarse, ya que y^* no se puede medir. La variable que sí se observa es la variable dicotómica y , que está relacionada con la variable latente de la siguiente manera:

$$y = \begin{cases} 1 & \text{si } y^* > 0 \\ 0 & \text{si } y^* \leq 0 \end{cases} \quad (4)$$

Suponiendo que el término de error e es independiente de \mathbf{x} y que se distribuye de forma simétrica alrededor de cero, a partir de (3) y (4) se deriva la probabilidad de respuesta para la variable observada y

$$P(y = 1 | \mathbf{x}) = P(y^* > 0 | \mathbf{x}) = P[e > -(\beta_0 + \mathbf{x}\boldsymbol{\beta}) | \mathbf{x}] = 1 - G[-(\beta_0 + \mathbf{x}\boldsymbol{\beta})] = G(\beta_0 + \mathbf{x}\boldsymbol{\beta}) \quad (5)$$

Modelo de Regresión Logística

Se comienzan a utilizar para la puntuación de riesgo de crédito a partir de que Wiginton (**1980**) mostrase su superioridad sobre los modelos discriminantes de clasificación utilizados por entonces (modelos que sólo permitían clasificación, no puntuación individual, identificaban a cada individuo dentro de uno de los dos grupos: buenos y malos)

$$P(y=1|x_1, x_2, x_3, \dots, x_m) = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m)}}$$

$\beta_0, \beta_1 \dots \beta_m$ \longrightarrow Parámetros a estimar

$x_1, x_2, x_3, \dots, x_m$ \longrightarrow Variables Explicativas

$P(y=1 | \mathbf{x})$ \longrightarrow Variable Objetivo o Vble respuesta: toma valores entre 0 y 1

Estimación por Máxima Verosimilitud de las betas

La estimación puede hacerse por pasos o con penalizaciones si se quiere para evitar problemas de multicolinealidad y seleccionar las variables que finalmente se utilizarán para predecir la probabilidad de impago (o el logaritmo del Odd Ratio)

¿Por qué se dice que la logística es un modelo “lineal”?

$$P(\text{loan_status} = 1 \mid x_1, \dots, x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m)}} = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m}}$$

$$P(\text{loan_status} = 0 \mid x_1, \dots, x_m) = 1 - \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m}} = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m}}$$

$$\frac{P(\text{loan_status} = 1 \mid x_1, \dots, x_m)}{P(\text{loan_status} = 0 \mid x_1, \dots, x_m)} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m} \rightarrow \text{odds in favor of loan_status=1}$$

$$\text{Logit}(P) = \ln(P/(1-P)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

Diagnosis del Modelo

$$P(y = 1 \mid \mathbf{x}) = P(y = 1 \mid x_1, x_2, \dots, x_k) = G(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$$

Efectos marginales

$$\frac{\partial p(\mathbf{x})}{\partial x_j} = g(\beta_0 + \mathbf{x}\boldsymbol{\beta}) \beta_j$$

Efectos marginales regresión logística

$$\frac{P(\text{loan_status} = 1 \mid x_1, \dots, x_m)}{P(\text{loan_status} = 0 \mid x_1, \dots, x_m)} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_m x_m}$$

Las betas no se pueden interpretar más que con el signo

$$\text{Logit}(P) = \ln(P/(1-P)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

If variable x_j goes up by 1 \rightarrow The odds are multiplied by e^{β_j}

$\beta_j < 0 \rightarrow e^{\beta_j} < 1 \rightarrow$ The odds decrease as x_j increases

$\beta_j > 0 \rightarrow e^{\beta_j} > 1 \rightarrow$ The odds increase as x_j increases

Medidas de la Bondad global del ajuste

No se puede calcular el R^2

En los modelos de regresión de probabilidad **no existen residuos** sobre las probabilidades estimadas. Es decir, **no se conoce la verdadera probabilidad que a priori tiene cada cliente de ser mal pagador**. Sí se conoce, por el contrario, su probabilidad a posteriori, esto es, sí se conoce **si al final** (en la ventana de observación) **cada uno de los clientes incurrieron en impago o no**.

➔ **Pseudo R^2** MacFadden **pseudo $R^2 = 1 - (L_{NR}/L_0)$**

L_{NR} es la función de log-verosimilitud para el modelo estimado y L_0 es la correspondiente al modelo con sólo el término constante

➔ **Significación Conjunta:**
Estadístico Razón de Verosimilitudes $LR = -2 \cdot [L_{NR} - L_0] \sim \text{Chi}^2 \text{ (g.d.} = k-1 \text{)}$

➔ **Test Hosmer-Lemeshow (H-L).** *Hipótesis nula: el modelo está bien especificado*

Comprueba si los porcentajes de $y=1$ e $y=0$ observados y los esperados o predichos por el modelo son similares (entre los 10 deciles de la predicción $P(y=1)$ del modelo) y se distribuye como Chi^2 (g.d. = $Q-2=8$)



Medidas de bondad global por capacidad predictiva

En los modelos de regresión de probabilidad **no existen residuos** sobre las probabilidades estimadas. Es decir, no se conoce la verdadera probabilidad que a priori tiene cada cliente de ser mal pagador. Sí se conoce, por el contrario, su probabilidad a posteriori, esto es, sí se conoce si al final (en la ventana de observación) cada uno de los clientes incurrieron en impago o no.

→ Matrices de Confusión

Aunque no existan residuos como tales, sí **puedo conocer cual es la capacidad predictiva del modelo**. Con los modelos de probabilidad somos capaces de estimar una probabilidad de impago para cada cliente. **A partir de dicha probabilidad de impago estimada por el modelo podemos realizar un pronóstico**, a priori, **sobre si un determinado cliente va a incurrir en impago o no**.

Comparando esos pronósticos realizados a priori sobre si un cliente incurrirá en impago, con los datos observados de dicho cliente (si incurrió realmente en impago o no en la ventana de observación), **es posible construir una matriz de pronósticos correctos e incorrectos** (denominada matriz de confusión, **y a partir de ahí elaborar diferentes indicadores de lo bien o mal que predice el modelo**).

Modelo de probabilidad

$$P(y = 1 \mid \mathbf{x}) = P(y = 1 \mid x_1, x_2, \dots, x_k) = G(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k) = G(\beta_0 + \mathbf{x}\boldsymbol{\beta})$$

Estimación del modelo

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$$

A partir de aquí, **para cada individuo** i se obtiene una **probabilidad de impago** según el modelo estimado

$$\hat{P}(y_i = 1 \mid x_{i1}, x_{i2}, \dots, x_{ik}) = G(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}) = G(\hat{\beta}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}})$$

Pronóstico

$$\begin{cases} \text{si } G(\hat{\beta}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}) > c \rightarrow \text{la predicción de } y_i \text{ es } 1 \\ \text{si } G(\hat{\beta}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}) \leq c \rightarrow \text{la predicción de } y_i \text{ es } 0 \end{cases}$$

El punto de corte c para la probabilidad estimada resulta crucial para hacer los pronósticos.

La probabilidad estimada por el modelo será siempre la misma, por ejemplo dada una estimación de las $\hat{P}_i = 0.25$:

- con un $c=0.5$ el pronóstico será de **no impago** $\hat{y}_i = 0$
- con un $c=0.15$ el pronóstico será de **impago** $\hat{y}_i = 1$

La elección del punto de corte de la probabilidad estimada de impago para realizar los pronósticos es una pieza clave de los modelos de puntuación de riesgo de crédito



Id cliente	Probabilidad Estimada de impago	\hat{y} Pronóstico de <i>impago</i> (con $c=0.5$)	Pronóstico		y <i>Impago</i> Observado real	<i>¿Cómo ha sido el pronóstico?</i>	Tipo de pronóstico
			$\hat{y} = 0 \rightarrow$ <i>Negativo</i>	$\hat{y} = 1 \rightarrow$ <i>Positivo</i>			
1	0.25	0	Negativo		0	Acierto	Verdadero Negativo
2	0.15	0	Negativo		0	Acierto	Verdadero Negativo
3	0.69	1		Positivo	0	Error	Falso Positivo
4	0.42	0	Negativo		1	Error	Falso Negativo
5	0.58	1		Positivo	1	Acierto	Verdadero Positivo
6	0.13	0	Negativo		1	Error	Falso Negativo
7	0.32	0	Negativo		0	Acierto	Verdadero Negativo
8	0.41	0	Negativo		0	Acierto	Verdadero Negativo
9	0.01	0	Negativo		0	Acierto	Verdadero Negativo
10	0.35	0	Negativo		1	Error	Falso Negativo
11	0.63	1		Positivo	0	Error	Falso Positivo
12	0.57	1		Positivo	1	Acierto	Verdadero Positivo
13	0.15	0	Negativo		0	Acierto	Verdadero Negativo
14	0.31	0	Negativo		0	Acierto	Verdadero Negativo
15	0.67	1		Positivo	0	Error	Falso Positivo
...

$\begin{cases} \text{si } G(\hat{\beta}_0 + \mathbf{x}_i \hat{\beta}) > c \rightarrow \text{la predicción de } y_i \text{ es } 1 \\ \text{si } G(\hat{\beta}_0 + \mathbf{x}_i \hat{\beta}) \leq c \rightarrow \text{la predicción de } y_i \text{ es } 0 \end{cases}$

Matriz de confusión

Estado Real de
crédito o Impago
observado (y)

		Pronóstico de <i>impago</i> \hat{y} (con $c=0.5$)	
		NO impago $\hat{y} = 0$	IMPAGO $\hat{y} = 1$
Estado Real de crédito o Impago observado (y)	NO impago $y = 0$	Verdadero Negativo TN=7	Falso Positivo FP=3
	IMPAGO $y = 1$	Falso Negativo FN=3	Verdadero Positivo TP=2

Medidas de Bondad del Modelo basados en la Matriz de Confusión

Exactitud (*accuracy*) = $(TP + TN) / (TP + FP + TN + FN)$

Porcentaje total de aciertos (cuanto de bien pronosticamos y nos acercamos al verdadero valor)

Precisión (*precisión*) Porcentaje de **aciertos** dentro de los **pronósticos** de positivos o de negativos

Precisión (sobre positivos) = $TP / (TP + FP)$

Precisión (sobre negativos) = $TN / (TN + FN)$

→ **Tasa de fallos sobre los negativos**

Bad rate: $(1 - \text{Precision}(0)) = FN / (TN + FN)$

Indica los **Falsos Negativos**, es decir los préstamos aceptados pero que resultaron ser malos (impagados), sobre el total de préstamos aceptados (pronosticados como negativos)

$y = 0$: Negativo

$y = 1$: Positivo

Pronóstico \hat{y} dado un punto de corte c

	Pronóstico \hat{y} dado un punto de corte c	
	Pronóstico NO impago $\hat{y} = 0$	Pronóstico Sí impago $\hat{y} = 1$
Observado NO impago $y = 0$	Verdadero Negativo TN	Falso Positivo FP
Observado Sí impago $y = 1$	Falso Negativo FN	Verdadero Positivo TP

$\text{precisión}(0)$

$\text{precisión}(1)$

Error tipo II

1- potencia

Error tipo I

significatividad

Medidas de Bondad del Modelo basados en la Matriz de Confusión

Exactitud (*accuracy*) = $(TP + TN) / (TP + FP + TN + FN)$

Porcentaje total de aciertos (cuanto de bien pronosticamos y nos acercamos al verdadero valor)

$y=0$: Negativo

$y=1$: Positivo

Pronóstico \hat{y} dado un punto de corte c

	Pronóstico \hat{y} dado un punto de corte c	
	Pronóstico NO impago $\hat{y} = 0$	Pronóstico Sí impago $\hat{y} = 1$
Observado NO impago $y = 0$	Verdadero Negativo TN	Falso Positivo FP
Observado Sí impago $y = 1$	Falso Negativo FN	Verdadero Positivo TP
	\downarrow <i>precisión(0)</i>	\downarrow <i>precisión(1)</i>

\rightarrow *recall(0)*

\rightarrow *recall(1)*

Precisión (*precisión*) Porcentaje de **aciertos** dentro de los **pronósticos** de positivos o de negativos

Precisión (sobre positivos) = $TP / (TP + FP)$

Precisión (sobre negativos) = $TN / (TN + FN)$

Exhaustividad (*recall*) Porcentaje de pronósticos acertados dentro del total **observado** de positivos o de negativos reales

Exhaustividad-*recall* (sobre positivos) = $TP / (FN + TP)$

Exhaustividad-*recall* (sobre negativos) = $TN / (TN + FP)$

El total de impagos, cuántos he acertado

Sensibilidad (*Sensitivity*): *exhaustividad (recall) sobre positivos*, indica cuanto de los verdaderos malos clientes (positivos) somos capaces de detectar (% de positivos correctamente identificados)

Especificidad (*Specificity*): *exhaustividad (recall) sobre negativos* indica cuanto de los verdaderos buenos clientes somos capaces de detectar (% de negativos correctamente identificados)

1- Especificidad (1- *Specificity*): $(1 - \text{exhaustividad (recall) sobre negativos}) = FP / (TN + FP)$ indica el porcentaje de fallos sobre los negativos, cuanto de los verdaderos buenos clientes pronosticamos como malos (% de negativos erróneamente identificados como positivos)

Medidas de Bondad del Modelo basados en la Matriz de Confusión

Exactitud (*accuracy*) = $(TP + TN) / (TP + FP + TN + FN)$

Porcentaje total de aciertos (cuanto de bien pronosticamos y nos acercamos al verdadero valor)

$y=0$: Negativo

$y=1$: Positivo

Pronóstico \hat{y} dado un punto de corte c

	Pronóstico \hat{y} dado un punto de corte c	
	Pronóstico NO impago $\hat{y} = 0$	Pronóstico Sí impago $\hat{y} = 1$
Observado NO impago $y = 0$	Verdadero Negativo TN	Falso Positivo FP
Observado Sí impago $y = 1$	Falso Negativo FN	Verdadero Positivo TP
	↓ <i>precisión(0)</i>	↓ <i>precisión(1)</i>

→ *recall(0)*

→ *recall(1)*

Precisión (*precisión*) Porcentaje de **aciertos** dentro de los **pronósticos** de positivos o de negativos

Precisión (sobre positivos) = $TP / (TP + FP)$

Precisión (sobre negativos) = $TN / (TN + FN)$

Exhaustividad (*recall*) Porcentaje de pronósticos acertados dentro del total **observado** de positivos o de negativos reales

Exhaustividad-*recall* (sobre positivos) = $TP / (FN + TP)$

Exhaustividad-*recall* (sobre negativos) = $TN / (TN + FP)$

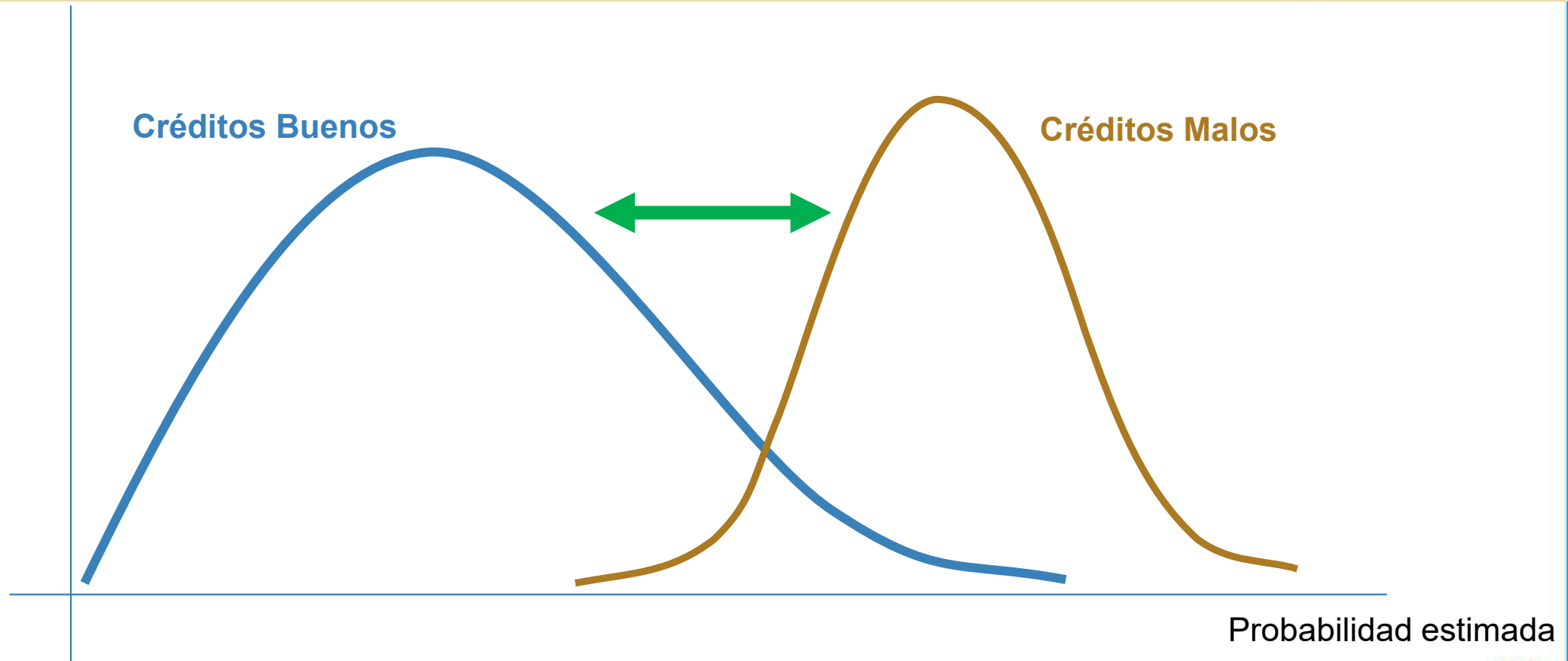
→ **Sensibilidad (*Sensitivity*)**: *exhaustividad (recall) sobre positivos*, indica cuanto de los verdaderos malos clientes (positivos) somos capaces de detectar (% de positivos correctamente identificados)

→ **Especificidad (*Specificity*)**: *exhaustividad (recall) sobre negativos* indica cuanto de los verdaderos buenos clientes somos capaces de detectar (% de negativos correctamente identificados)

→ **1- Especificidad (1- *Specificity*)**: $(1 - \text{exhaustividad (recall) sobre negativos}) = FP / (TN + FP)$ indica el porcentaje de fallos sobre los negativos, cuanto de los verdaderos buenos clientes pronosticamos como malos (% de negativos erróneamente identificados como positivos)

Evaluación del Modelo

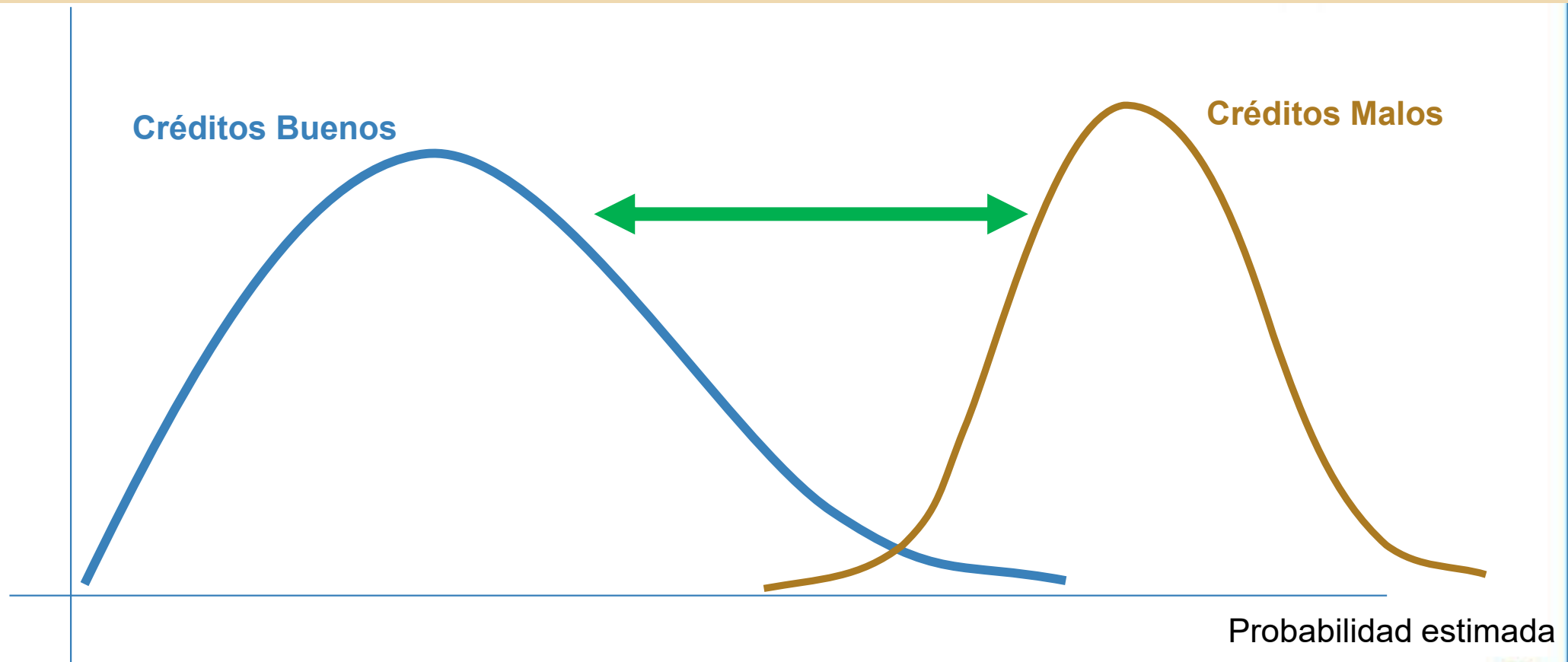
Objetivo: Queremos que nuestro modelo de estimación de riesgo de impago separe muy bien a los buenos de los malos créditos



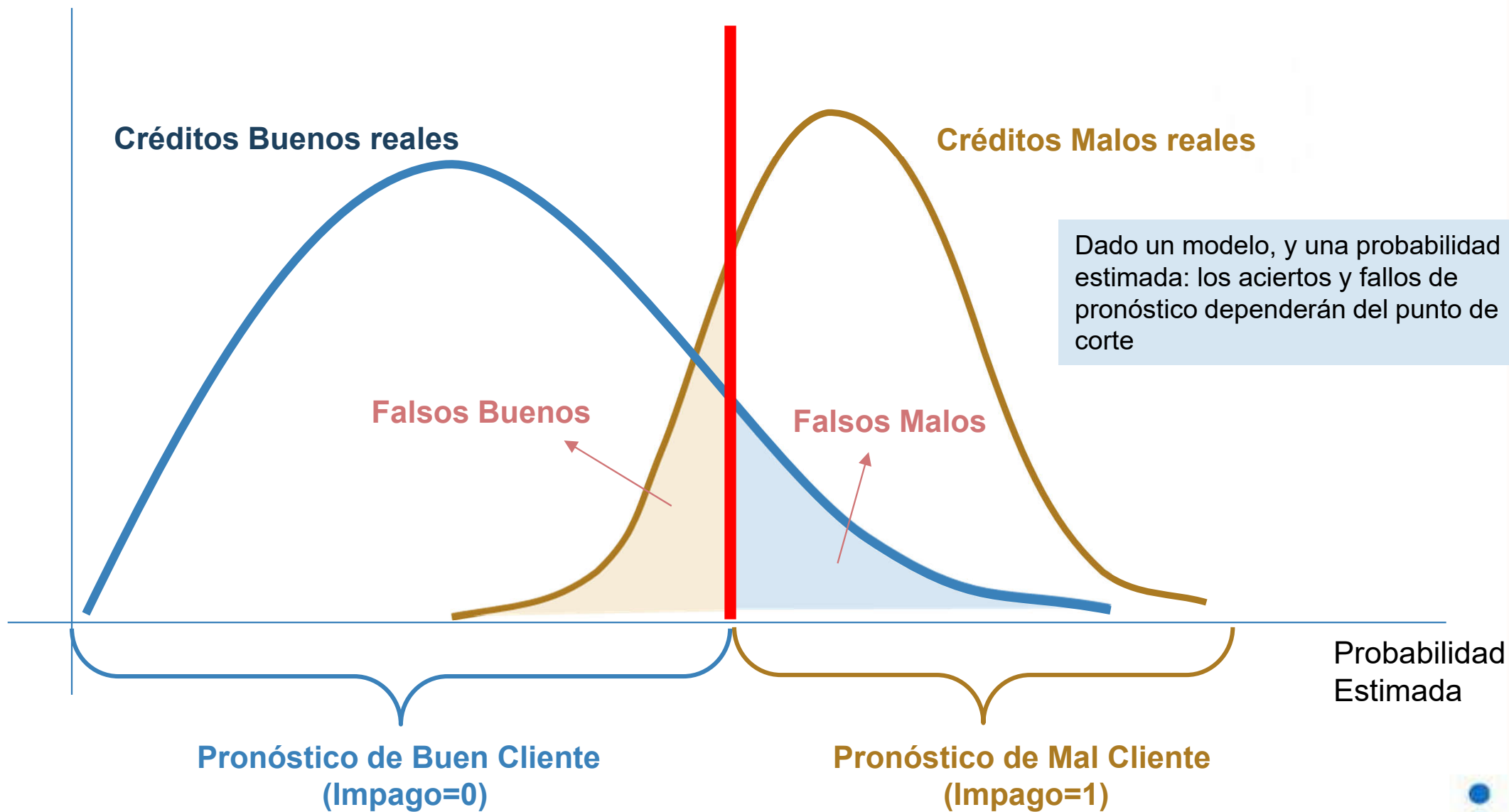
Estadístico de Divergencia: $D^2 = (\text{prob media de los buenos} - \text{prob media malos})^2 / \sigma^2$, con $\sigma^2 = (\sigma_G^2 + \sigma_B^2) / 2$

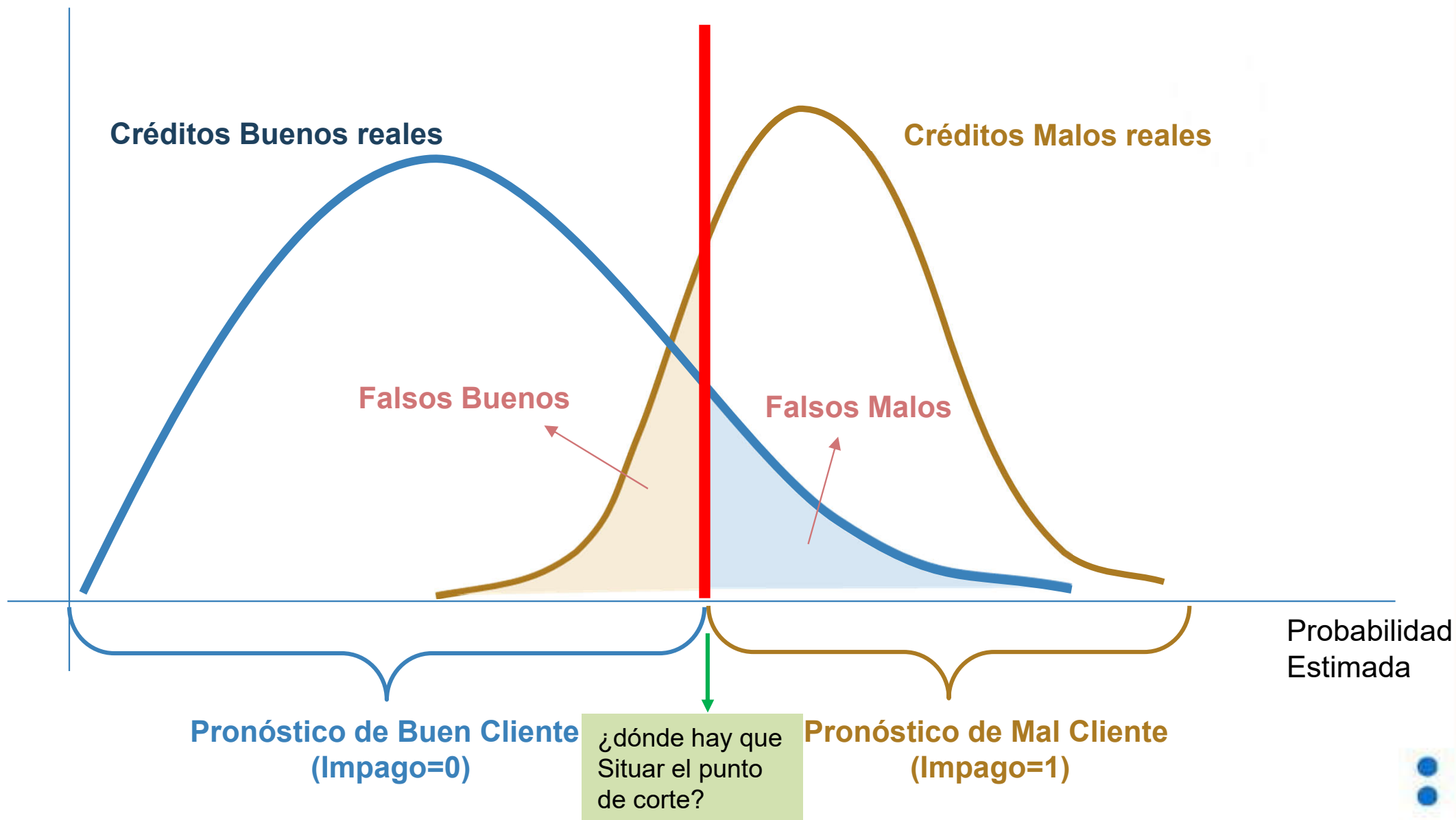
Evaluación del Modelo

Objetivo: Queremos que nuestro modelo de estimación de riesgo de impago separe muy bien a los buenos de los malos créditos



Estadístico de Divergencia: $D^2 = (\text{prob media de los buenos} - \text{prob media malos})^2 / \sigma^2$, con $\sigma^2 = (\sigma_G^2 + \sigma_B^2) / 2$





...y ¿dónde hay que poner el punto de corte?..... Siempre en $c=0.5$

$$\hat{P}(y_i = 1 \mid x_{i1}, x_{i2}, \dots, x_{ik}) = G(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}) = G(\hat{\beta}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}})$$

$$\text{Pronóstico} \begin{cases} \text{si } G(\hat{\beta}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}) > c & \text{la predicción de } y_i \text{ es 1} \\ \text{si } G(\hat{\beta}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}) \leq c & \text{la predicción de } y_i \text{ es 0} \end{cases}$$

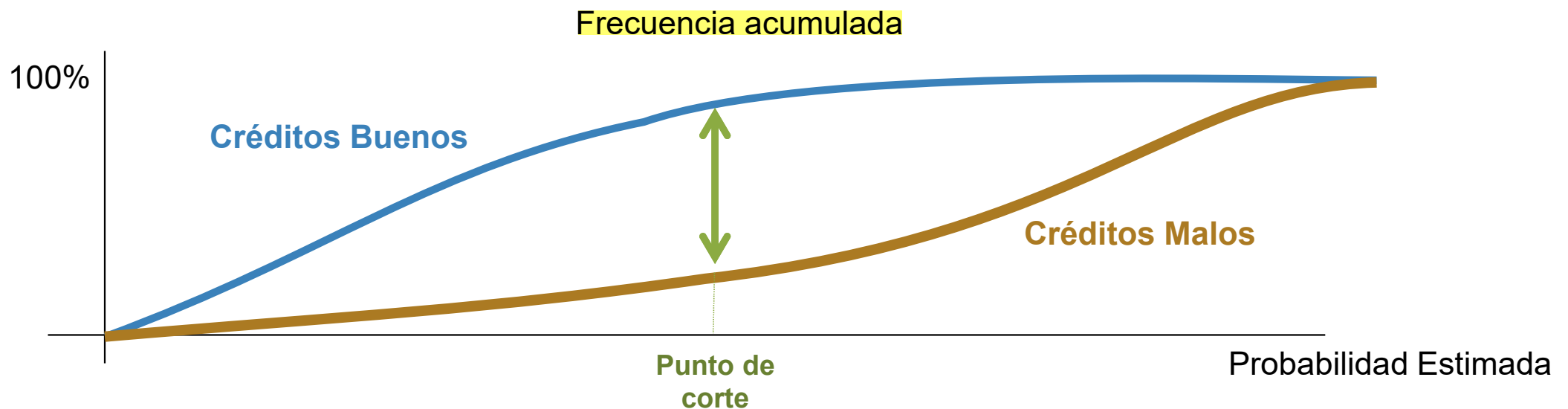
La elección del punto de corte de la probabilidad estimada de impago para realizar los pronósticos es una pieza clave de los modelos de puntuación de riesgo de crédito

El modelo de regresión logística reproduce o replica la frecuencia observada de $y=1$. La probabilidad media estimada por el modelo coincidirá con la frecuencia observada de $y=1$.

Por tal motivo el punto de corte debe situarse en torno a la frecuencia observada de $y=1$. (Cramer 1999, pag. 92), y todos los criterios de elección proporcionan puntos de corte en torno a esta frecuencia observada de $y=1$

Sólo debería utilizarse $c=0.5$ con muestras balanceadas (mismo número de $y=0$ y de $y=1$)

Estadístico KS: Máxima diferencia entre la distribución acumulada de los buenos y la distribución acumulada de los malos (se fija en la diferencia de toda la distribución y no sólo de la media)



KS: entre 0 y 100, (menor que 20 muy malo, mayor que 75 sospechosamente demasiado bueno)



Otro estadístico que suele utilizarse es el estadísticos **F₁Score**.

El **F₁-score** es una medida de la exactitud, o tasa de aciertos que combina tanto precisión como exhaustividad (*recall*). Es una media armónica de precisión y *recall*, para cada posible punto de corte, y que toma valores 0 cuando no hay ningún acierto en el pronóstico, y el valor 1 cuando se alcanza la máxima precisión y la máxima *recall*

$$F_1\text{-score} = \frac{precision \cdot recall}{\left(\frac{precision + recall}{2}\right)}$$

Como tanto precisión como exhaustividad (*recall*), pueden calcularse tanto para los positivos (malos clientes) como para los negativos (buenos clientes), suele utilizarse como medida **F₁-score** medio

$$F_1\text{-scoreMedio} = \frac{F_1score(y = 1) + F_1score(y = 0)}{2}$$

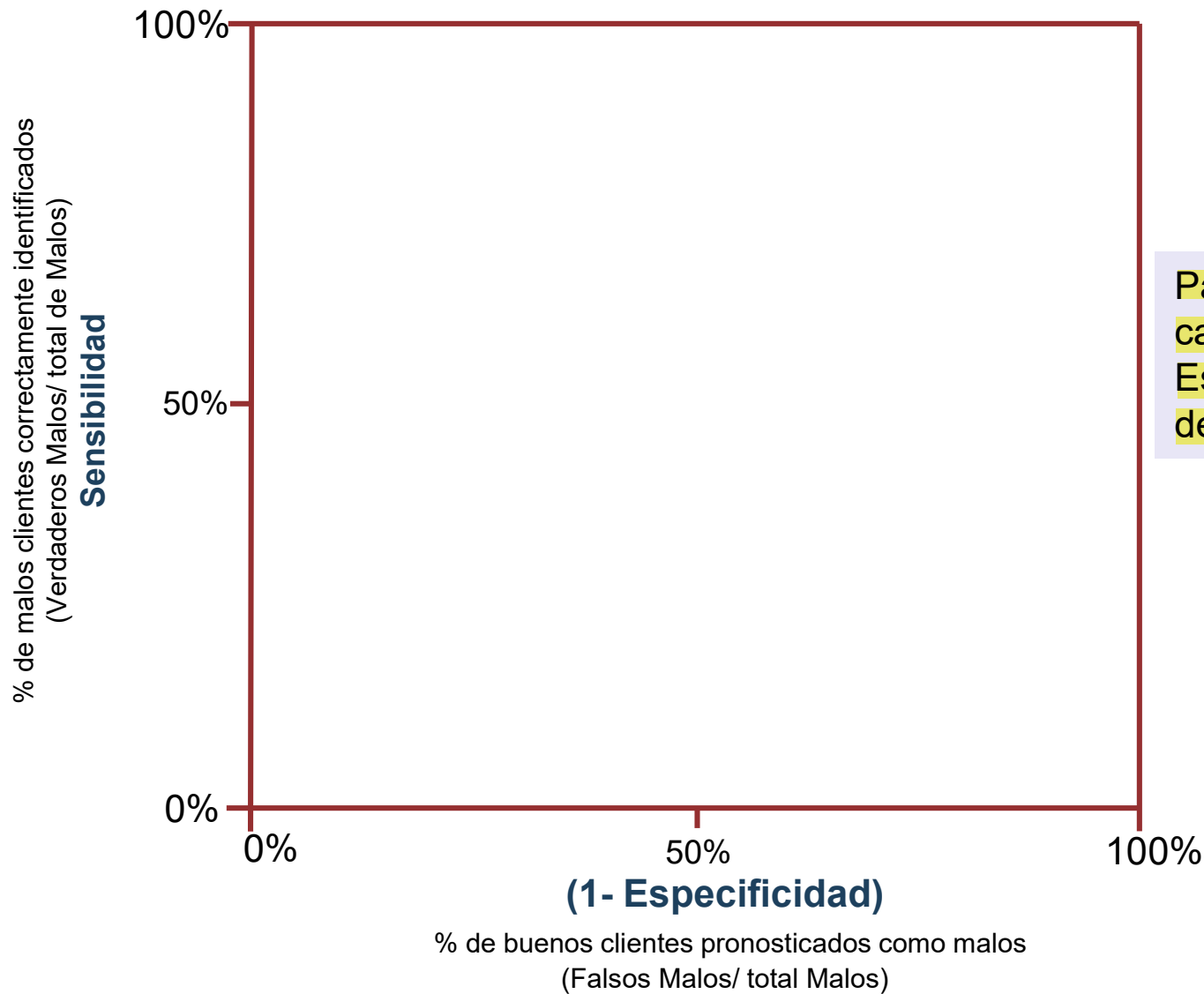
Una forma de encontrar el punto de corte óptimo, sería calculando el **F₁-score** para muchos puntos de corte, y seleccionar aquél que haga máximo el estadístico **F₁-score**

Existen medidas de bondad de ajuste (pronósticos correctos) basados en la matriz de confusión pero que **no dependen crucialmente de un único punto de corte** de probabilidad para realizar los pronósticos

Se calcula la Sensibilidad y la Especificidad **para muchos puntos de corte**

La Curva ROC, y el área por debajo de la Curva ROC

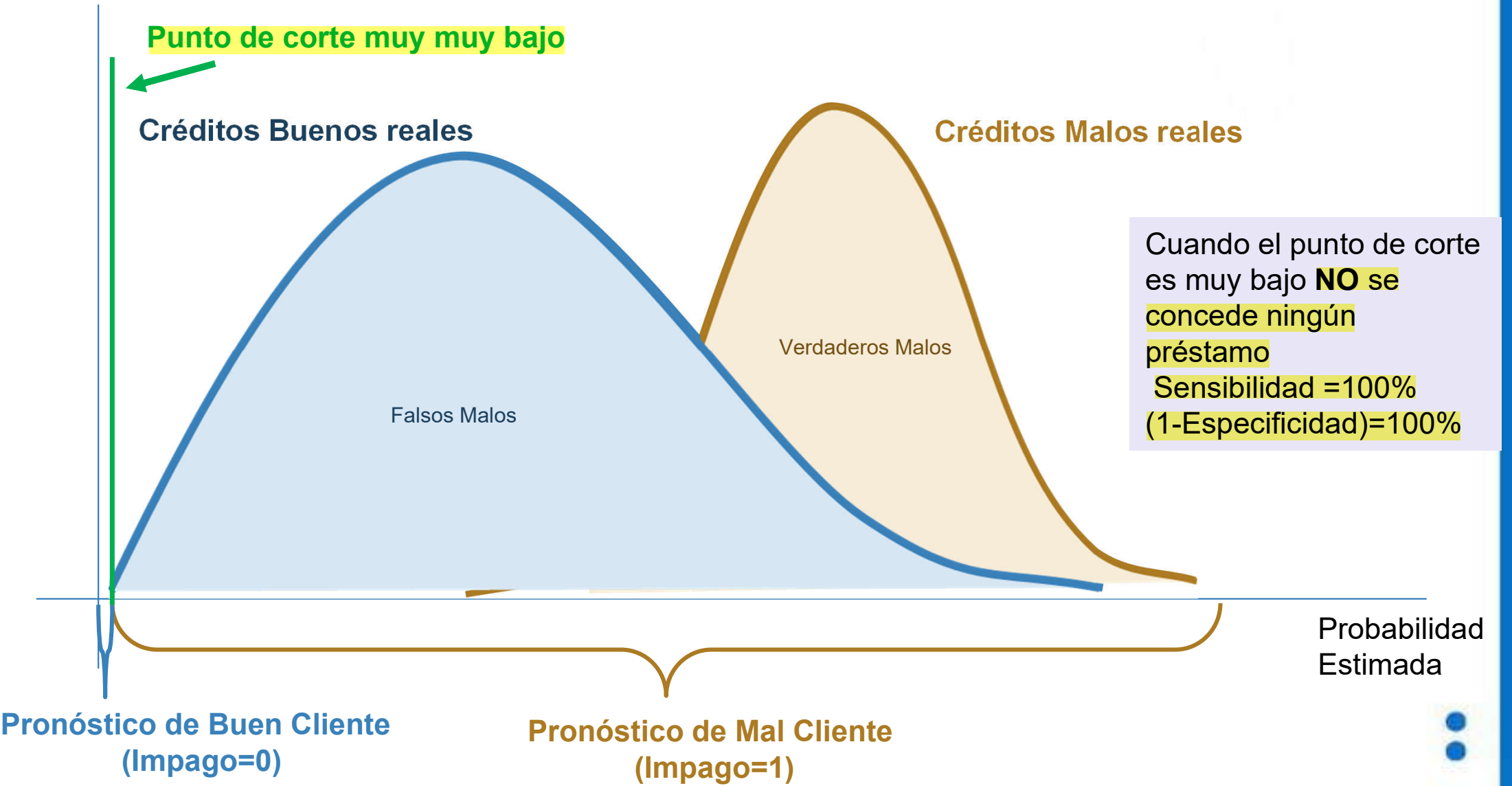
Curva ROC

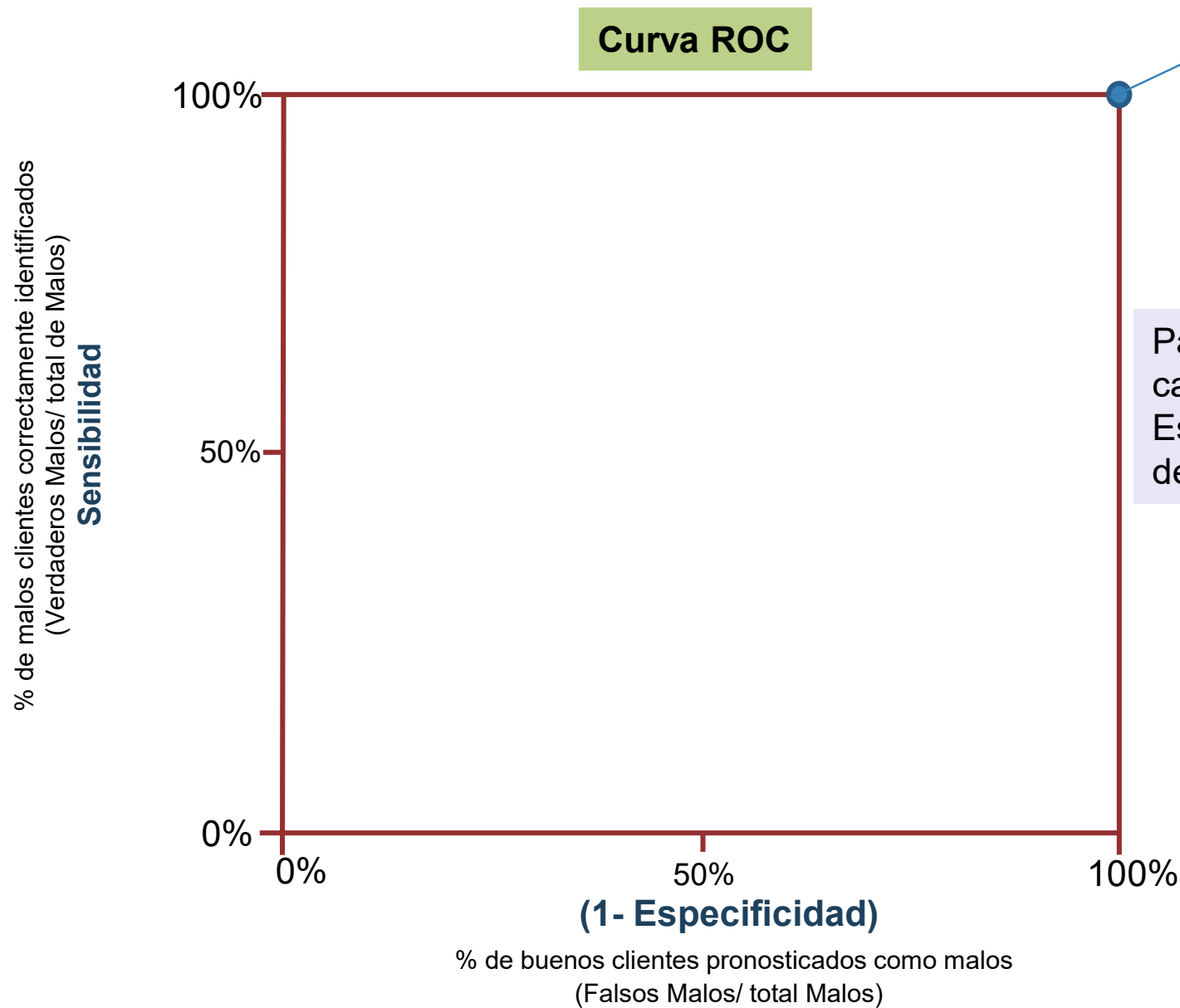


Para Construir la Curva ROC se calculan la sensibilidad y la Especificidad para diferentes valores del punto de corte

Sensibilidad (Sesibility): Verdaderos Malos/ total de Malos

(1- Especificidad) Falsos Malos/ total Malos





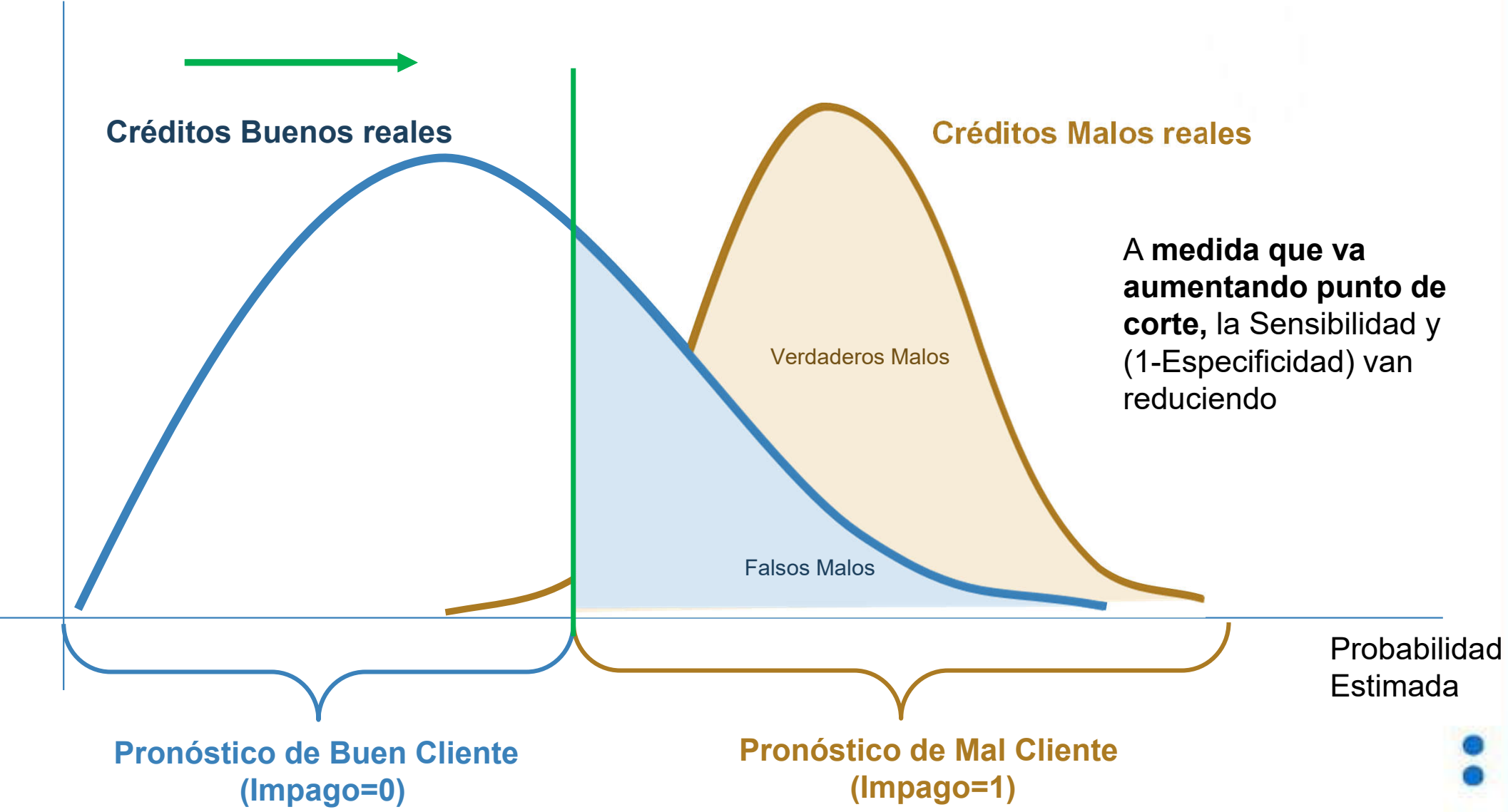
Cuando el punto de corte es muy muy bajo

Para Construir la Curva ROC se calculan la sensibilidad y la Especificidad para diferentes valores del punto de corte



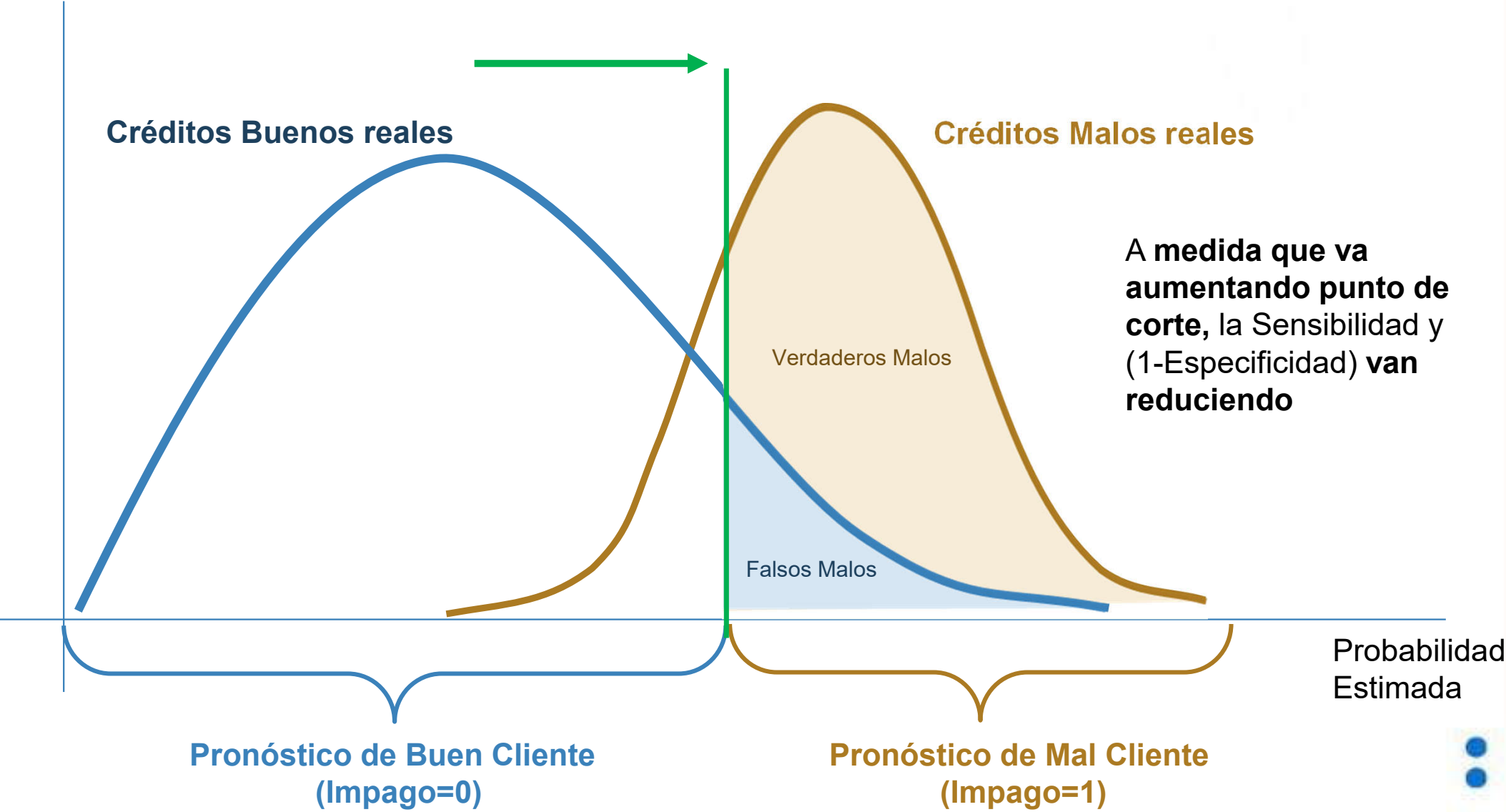
Sensibilidad (Sesibility): Verdaderos Malos/ total de Malos

(1- Especificidad) Falsos Malos/ total Malos



Sensibilidad (Sesibility): Verdaderos Malos/ total de Malos

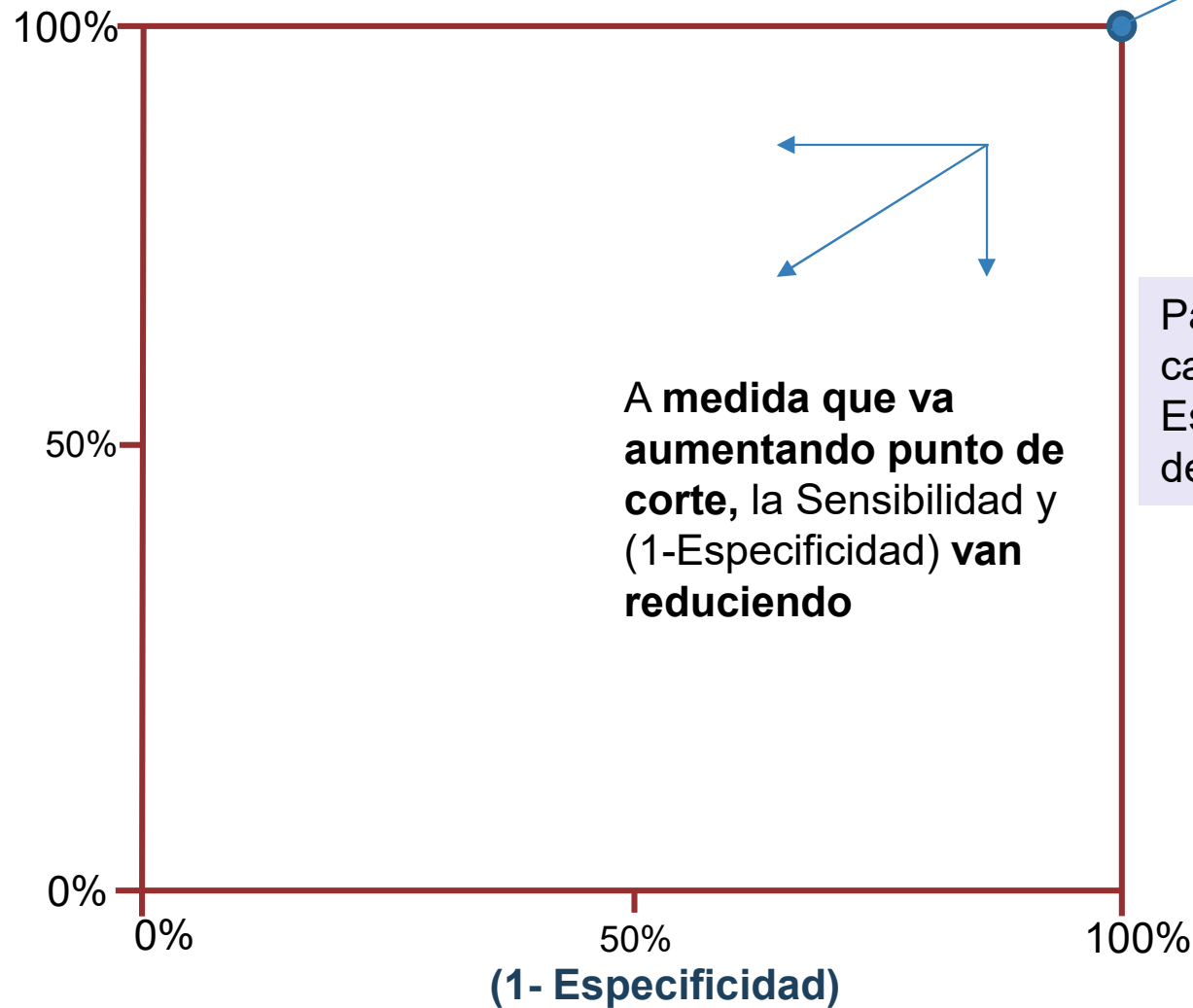
(1- Especificidad) Falsos Malos/ total Malos



% de malos clientes correctamente identificados
(Verdaderos Malos/ total de Malos)

Sensibilidad

Curva ROC



% de buenos clientes pronosticados como malos
(Falsos Malos/ total Malos)

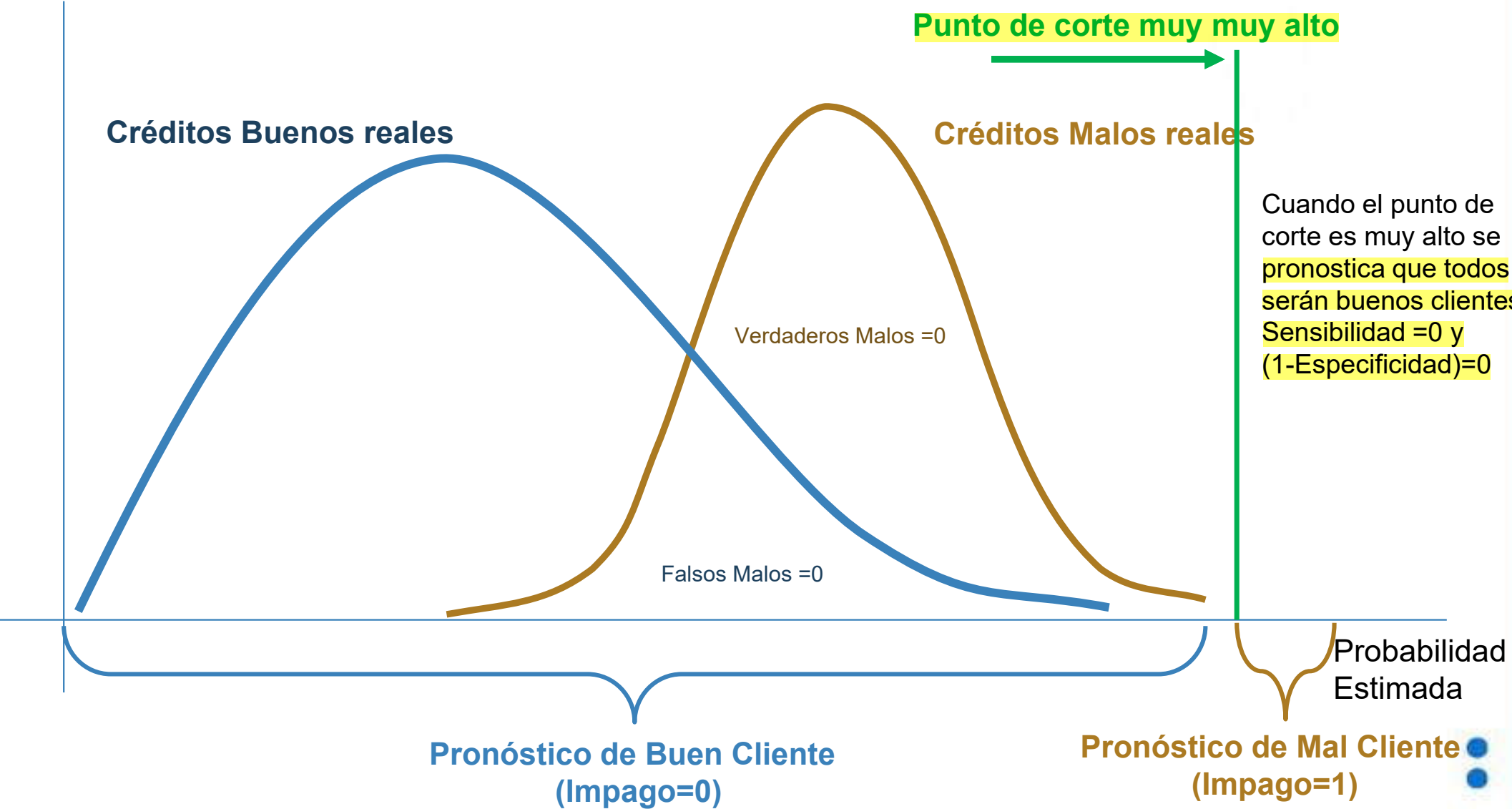
Cuando el punto de
corte es muy muy
bajo

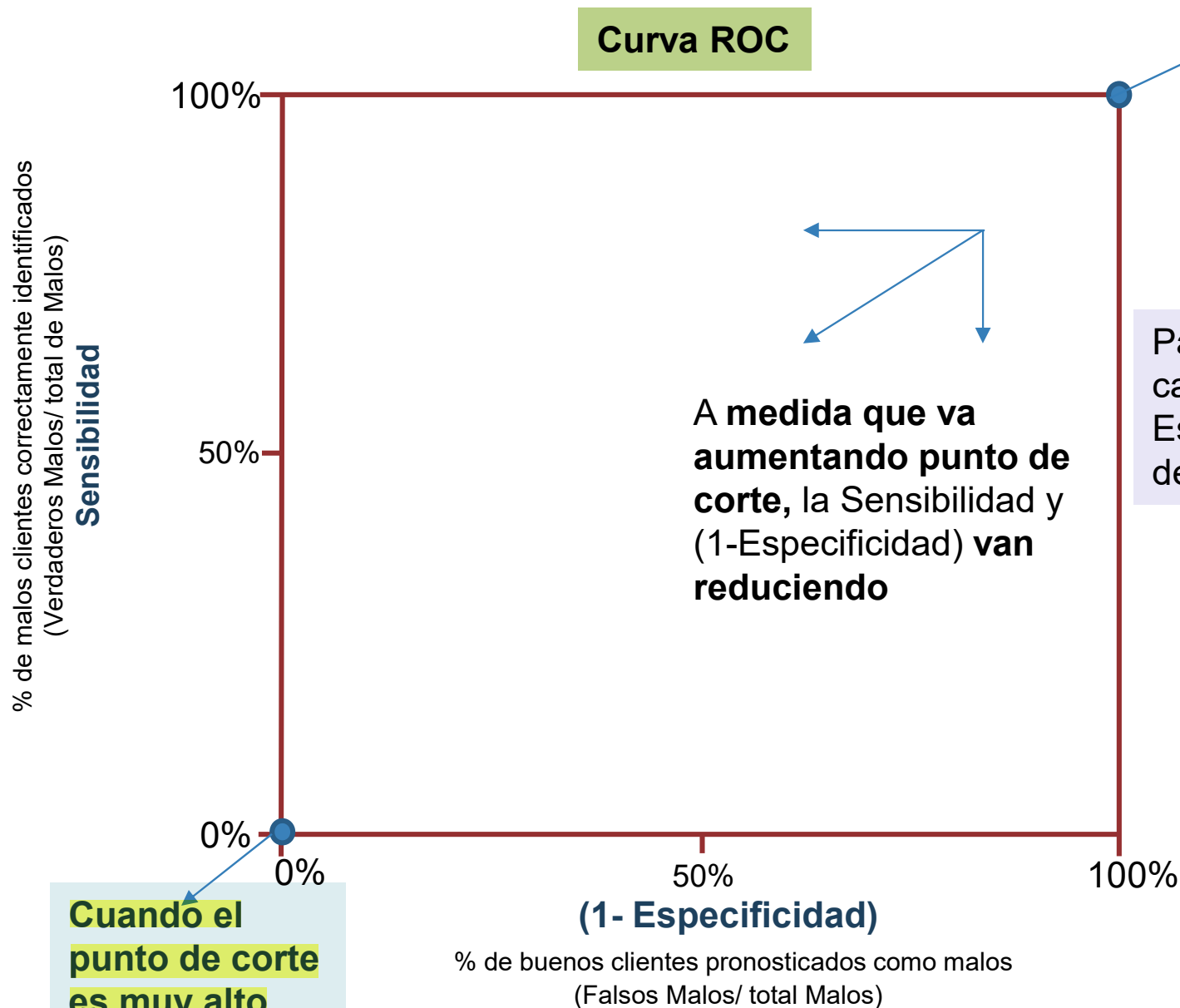
Para Construir la Curva ROC se
calculan la sensibilidad y la
Especificidad para diferentes valores
del punto de corte



Sensibilidad (Sesibility): Verdaderos Malos/ total de Malos

(1- Especificidad) Falsos Malos/ total Malos





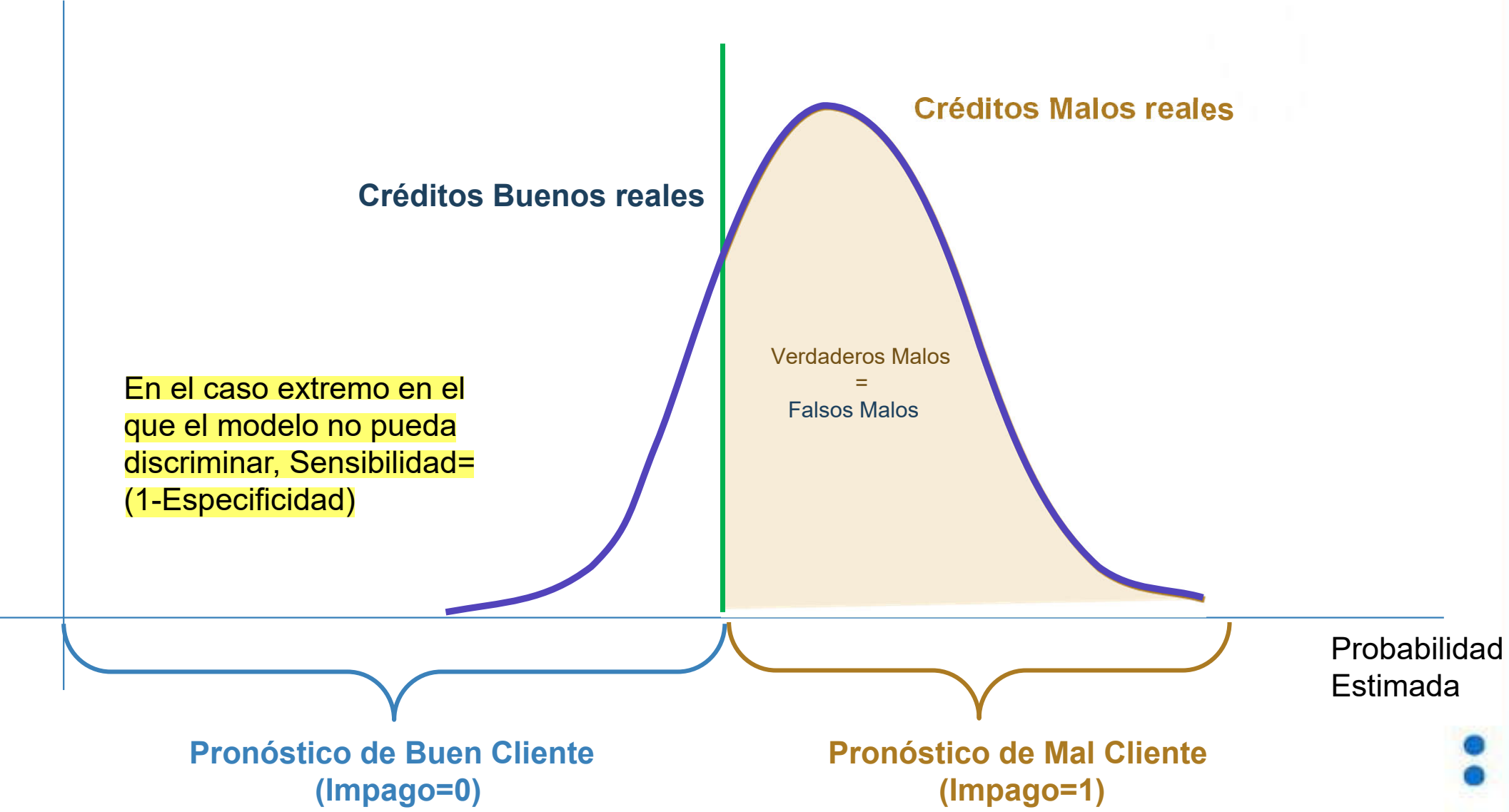
Cuando el punto de corte es muy bajo

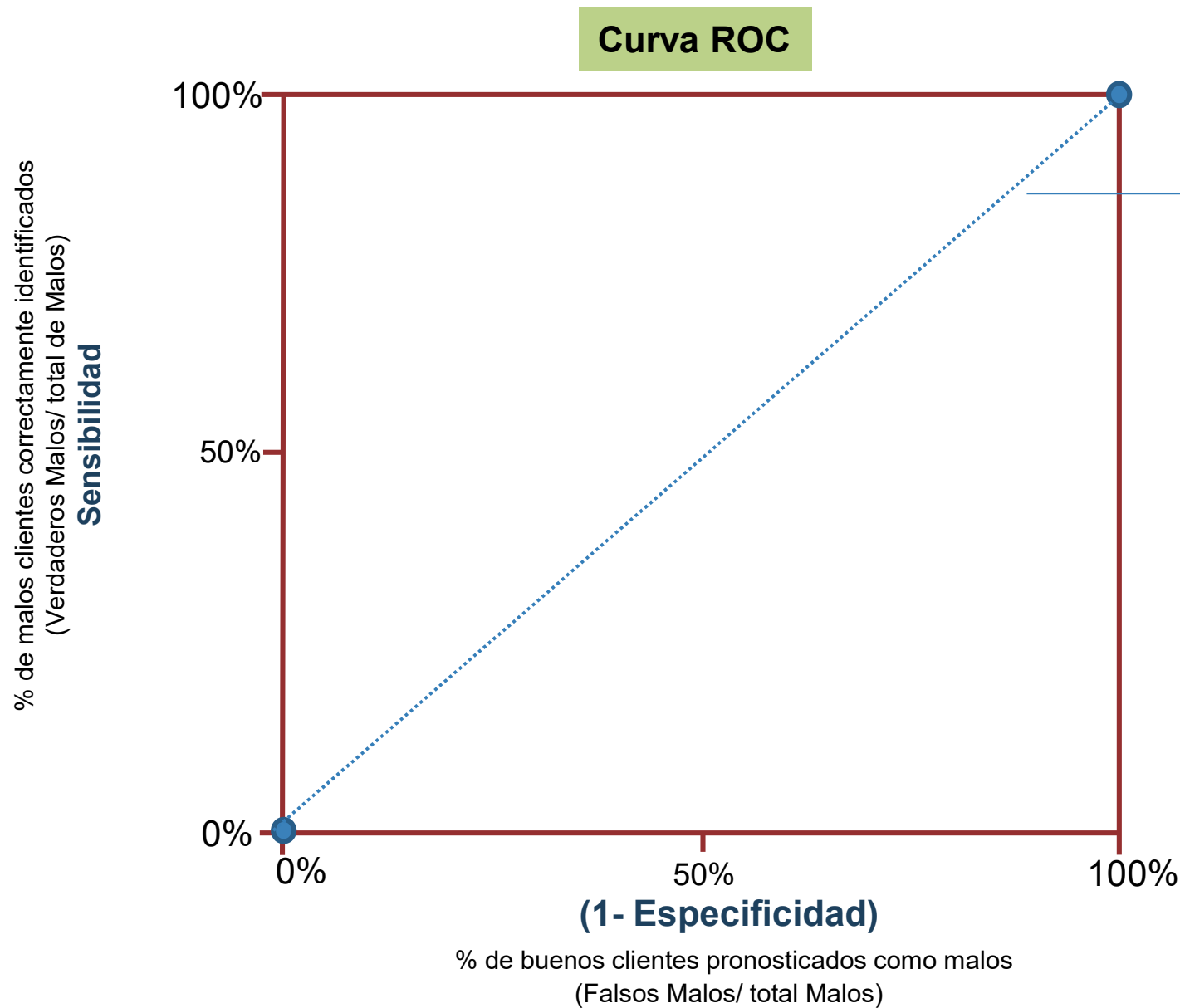
Para Construir la Curva ROC se calculan la sensibilidad y la Especificidad para diferentes valores del punto de corte

Cuando el punto de corte es muy alto

Sensibilidad (Sensitivity): Verdaderos Malos/ total de Malos

(1- Especificidad) Falsos Malos/ total Malos





Cuando la Curva ROC coincide con la bisectriz (recta 45°) el modelo carece de ninguna capacidad de discriminar a buenos y malos clientes (asignación de créditos puramente aleatoria)

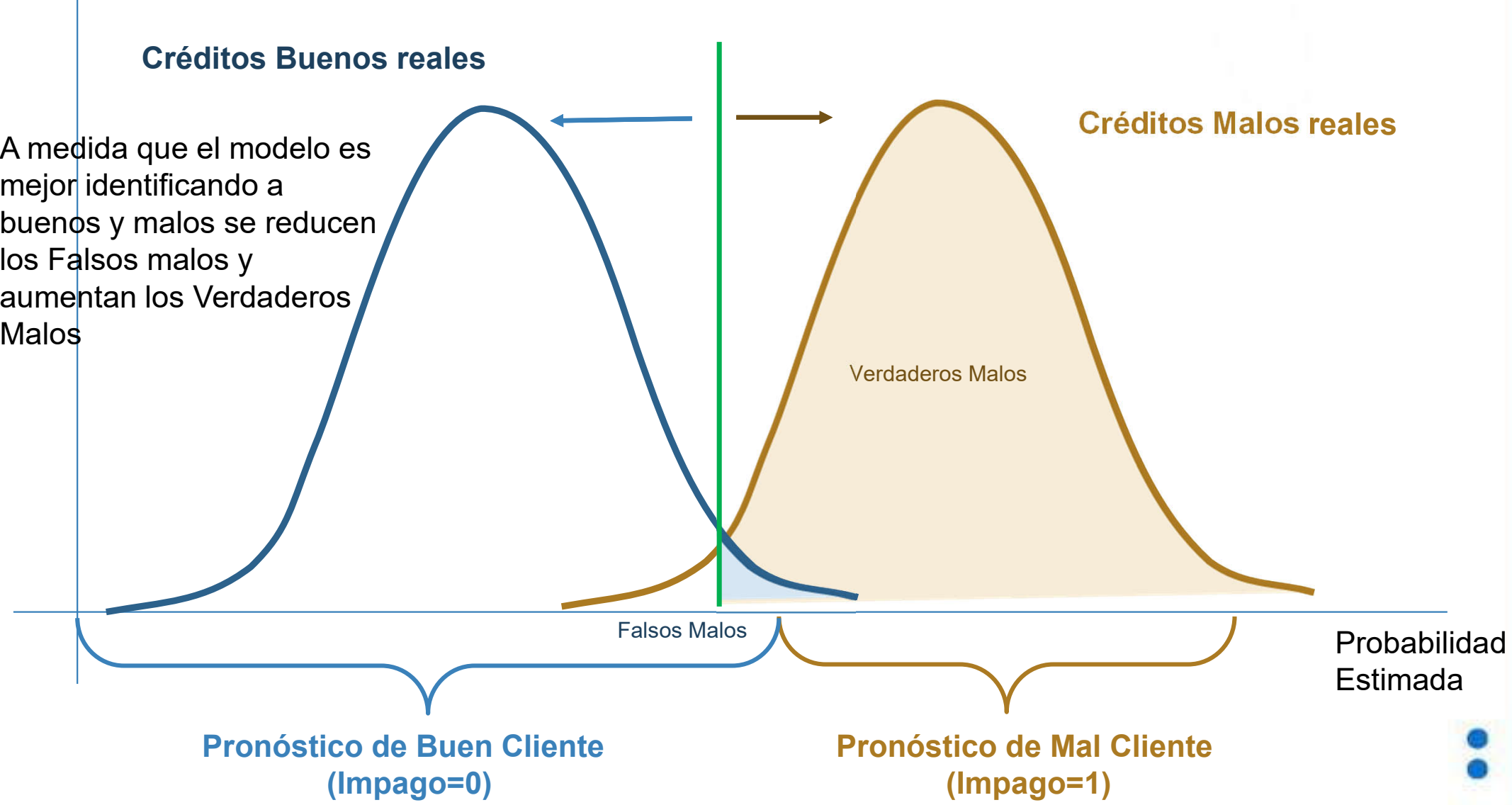
Sensibilidad (Sesibility): Verdaderos Malos/ total de Malos

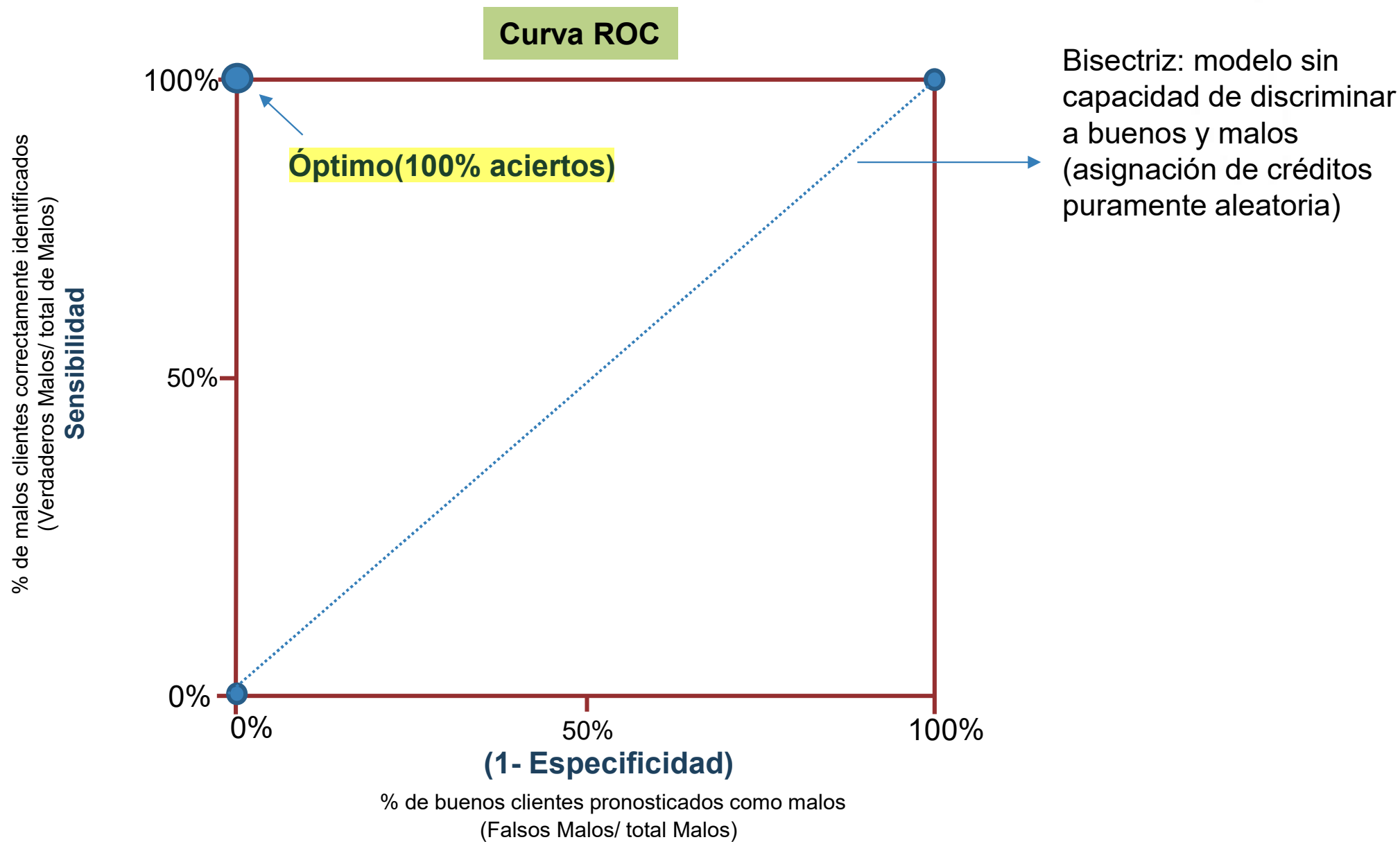
(1- Especificidad) Falsos Malos/ total Malos

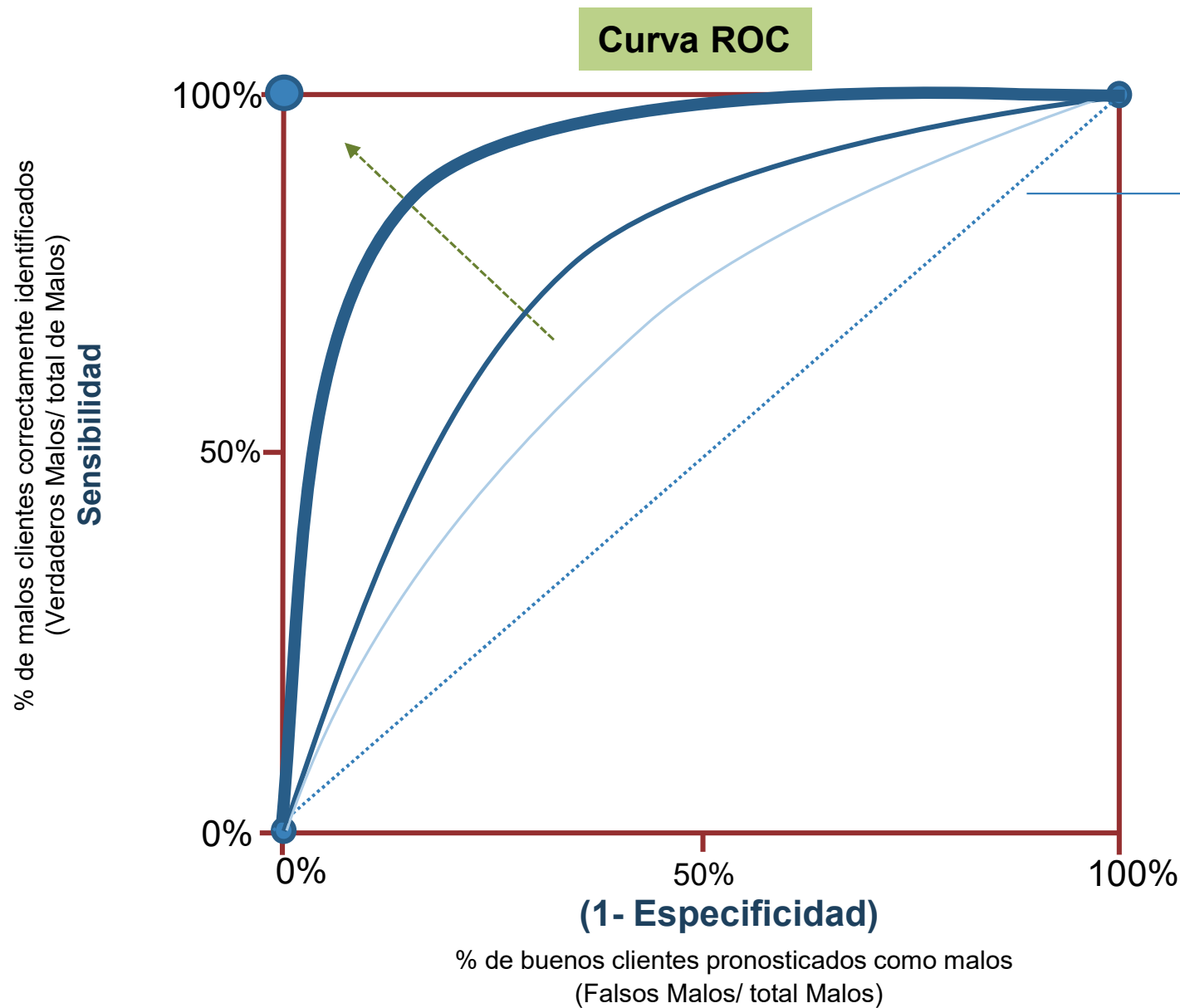


Sensibilidad (Sesibility): Verdaderos Malos/ total de Malos

(1- Especificidad) Falsos Malos/ total Malos







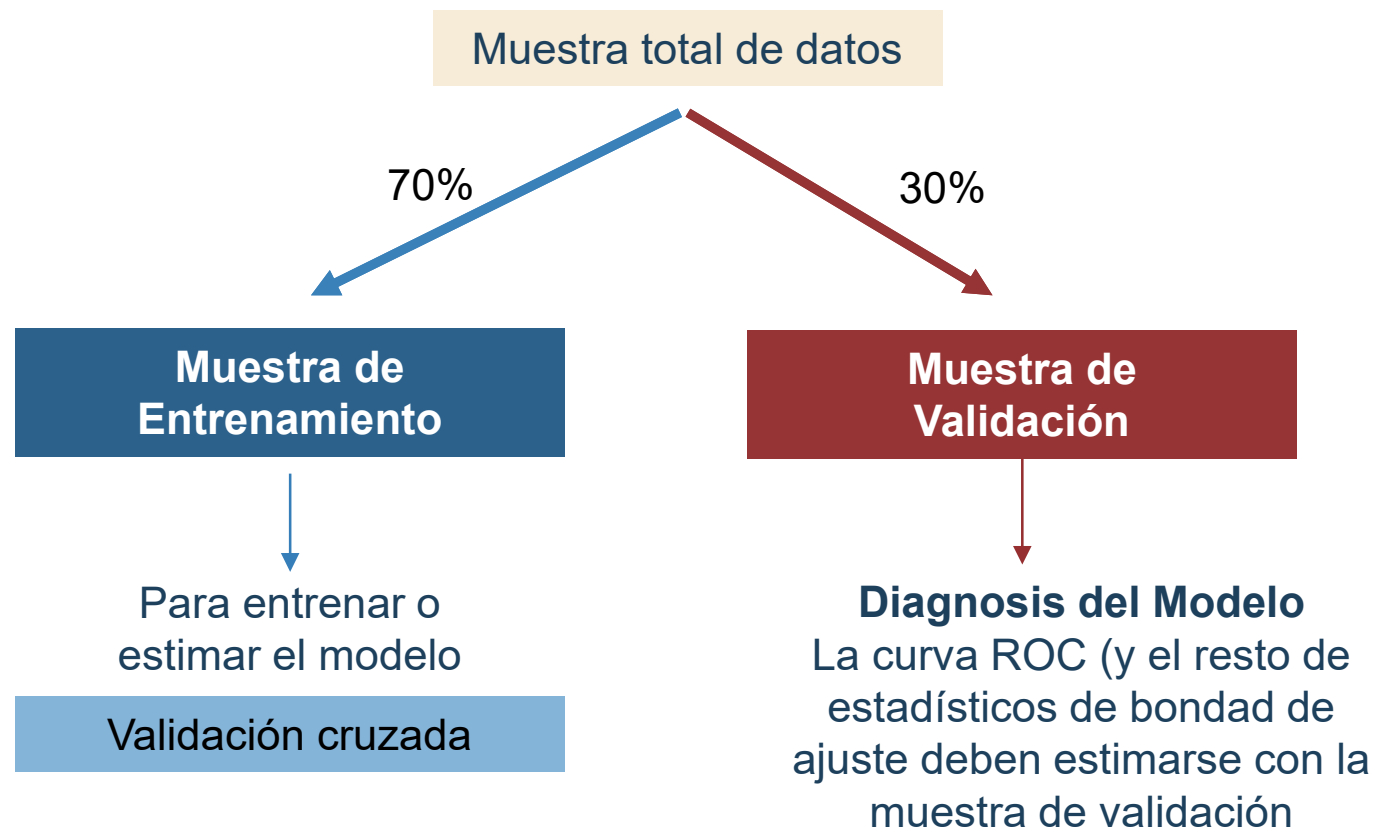
Bisectriz: modelo sin capacidad de discriminar a buenos y malos (asignación de créditos puramente aleatoria)

A Medida que la curva ROC se aleja de la Bisectriz y se acerca al punto óptimo la capacidad del modelo para discriminar a buenos y malos clientes va aumentando

Estadístico AUC: Área debajo de la curva ROC (en teoría de 0.5 a 1)
0.5-0.6:malo;
0.6-0.75:regular
0.75-1: aceptable

Una vez que se tiene el modelo:

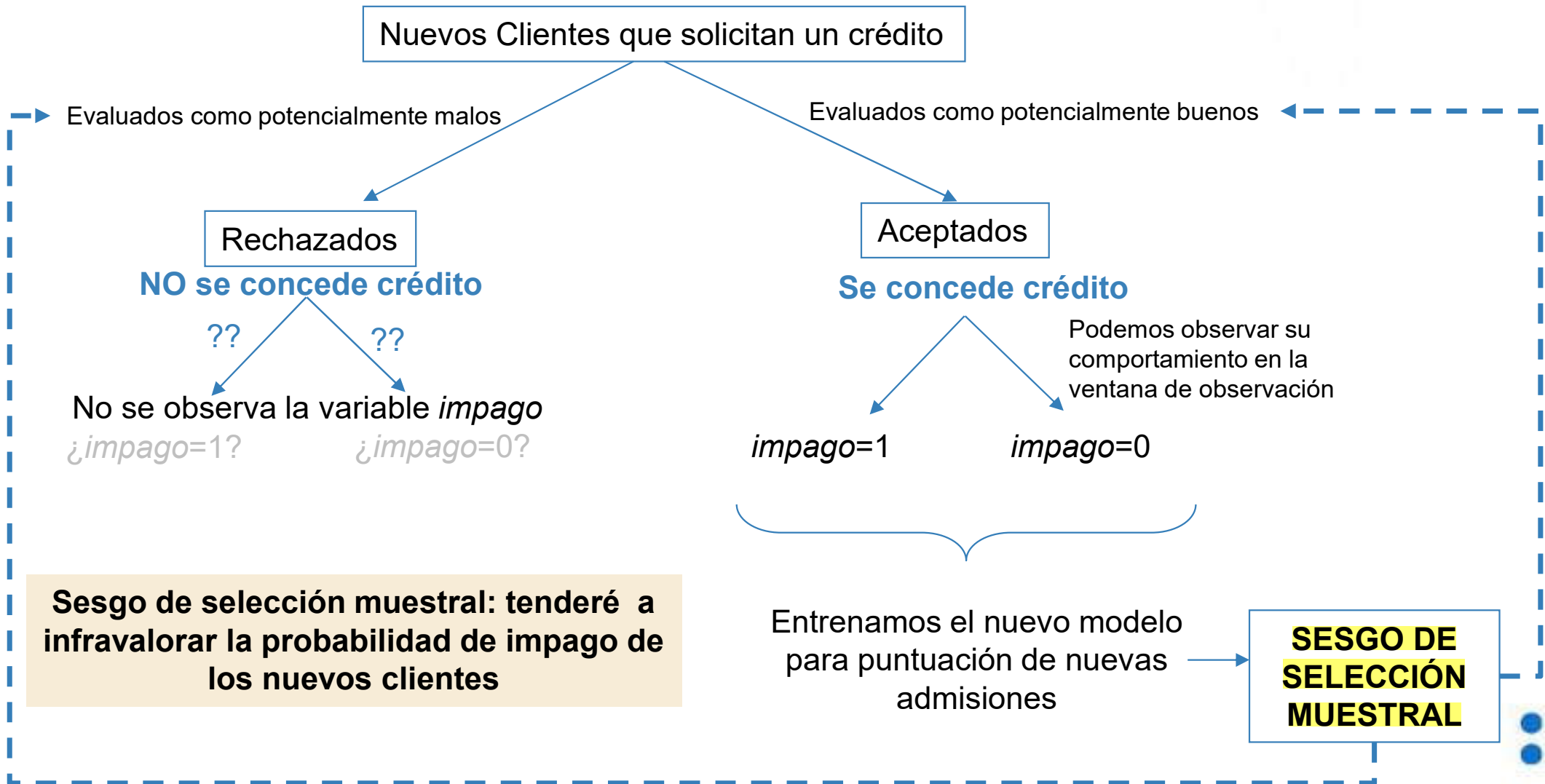
Diagnosis y Validación: siempre con la muestra de test (capacidad de predecir fuera de la muestra de entrenamiento para evitar el sobreajuste)



Como parte de la diagnosis debemos

- preguntarnos si el modelo tiene sentido
¿interpretación del modelo? (Caja negra?)
- Problema de variables omitidas
- Sesgos de selección muestral

Sesgo de Selección muestral en los modelos de scoring de riesgo



SOLUCION al problema del sesgo de selección: Inferencia de denegados

El problema es que sólo tenemos información sobre impagos de los clientes aceptados para concederles un crédito. También podemos tener algo de información de los clientes rechazados (para concederles el crédito), pero no sabemos si estos clientes rechazados han sido morosos o no (no tenemos información sobre la vble objetivo) porque como no le dimos el crédito nunca sabremos si hubieran sido buenos o malos

Sesgo de selección muestras: tenderé a infravalorar la probabilidad de impago de los nuevos clientes

Se han propuesto en la literatura dos SOLUCIONES al sesgo de selección muestral

1. Estimación en dos pasos de Heckman para la corrección del sesgo de selección

2. Inferencia de Denegados

- a. Asignar como malos (impagados) a todos rechazados
- b. Hard Cut-Off Augmentation
- c. Parceling Augmentation
- d. Fuzzy Augmentation

Requiere estimar un primer modelo sólo con los aceptados que posteriormente es utilizado para inferir que hubieran hecho los rechazados, y con esta inferencia volver a estimar un segundo modelo definitivo con aceptados y rechazados (inferidos) sin sesgo de selección

Nuevos Clientes que solicitan un crédito

→ Evaluados como potencialmente malos

Evaluados como potencialmente buenos ←

Rechazados

Aceptados

NO se concede crédito

Se concede crédito

??

??

No se observa la variable *impago*

Podemos observar su comportamiento en la ventana de observación

¿*impago*=1?

¿*impago*=0?

impago=1

impago=0

impago=1

impago=0

Inferencia de Denegados

Entrenamos un primer modelo para puntuar a los rechazados

Entrenamos un segundo modelo con aceptados y rechazados (inferidos) para puntuar a los nuevos clientes sin sesgos de selección muestral

Hard Cut-Off Augmentation

1. Build a scorecard model using the known good/bad population (that is, accepted applicants).
2. Score the rejected applicants with this model to obtain each rejected applicant's probability of default.
3. Create weighted cases for the rejected applicants.
4. Set a cut-off probability level above which an applicant is deemed *bad*. All applicants below this level are deemed *good*.
5. Add the inferred goods and bads back in with the known goods and bads and rebuild the scorecard.

Fuzzy Augmentation

1. Build a scorecard model using the known good/bad population (that is, accepted applicants) and score the rejected applicants with this model to obtain the probability of the rejected applicant being a good ($P(\text{good})$) and the probability of being a bad ($P(\text{bad})$).
2. Do not assign a reject to a good/bad class. Instead create two weighted cases for each rejected applicant using $P(\text{good})$ and $P(\text{bad})$.
3. Multiply $P(\text{good})$ and $P(\text{bad})$ by the user-specified rejection rate to form frequency variables.
4. Results are in two observations for the rejected applicants. One observation has a frequency variable (rejection weight $\times P(\text{good})$) and a target variable value of 0. The other observation has a frequency variable (rejection weight $\times P(\text{bad})$) and a target variable value of 1.

Parceling Augmentation

1. Build a scorecard model using the known good/bad population (that is, accepted applicants).
2. Score the rejected applicants with this model to obtain each rejected applicant's probability of default.
3. Create weighted cases for the rejected applicants.
4. The inferred good/bad status of the rejected applicants will be assigned randomly and proportional to the number of goods and bads in the accepted population **within each score range**.
5. If desired, apply the "event rate increase" factor to $P(\text{bad})$ to increase the proportion of bads in the rejects.
6. Add the inferred goods back in with the known goods and bads and rebuild the scorecard.

Score	# Bad	# Good	% Bad	% Good	Reject	Rej - Bad	Rej - Good
0-169	290	971	23.0%	77.0%	1548	379	1,267
170-179	530	2,414	18.0%	82.0%	1732	312	1,420
180-189	365	2,242	14.0%	86.0%	3719	521	3,198
190-199	131	1,179	10.0%	90.0%	7334	733	6,601
200-209	211	2,427	8.0%	92.0%	1176	94	1,082
210-219	213	4,047	5.0%	95.0%	3510	176	3,342
220-229	122	2,928	4.0%	96.0%	7211	288	6,923
230-239	139	6,811	2.0%	98.0%	3871	77	3,794
240-249	88	10,912	0.8%	99.2%	4773	38	4,735
250+	94	18,706	0.5%	99.5%	8962	45	8,937

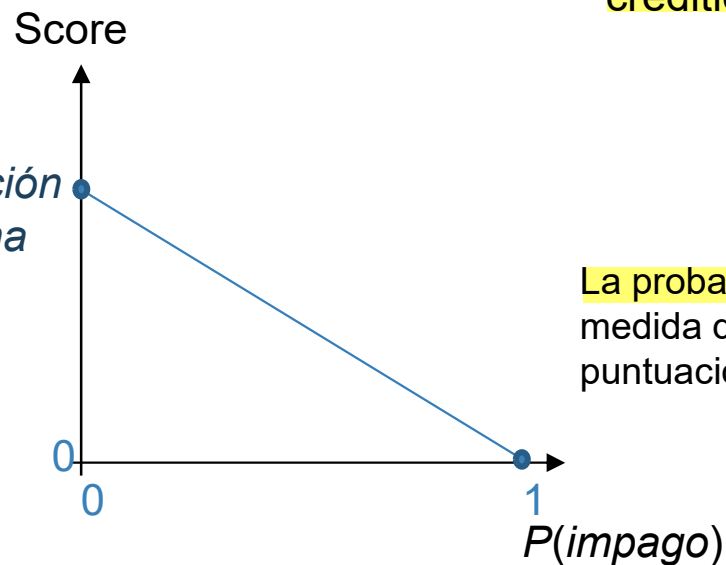
UNA VEZ que se le ha inferido a cada rechazado una variable objetivo se vuelve a rehacer todo el modelo (desde el principio)

¿Otra vez?..... Sí otra vez, porque todo lo que se hizo anteriormente sólo con los clientes aceptados tenía un sesgo: Así que hay que repetir todo desde la selección de variables.....

Una vez concluida la Diagnósis se pasa a la construcción de la **tarjeta de puntuación** y a la puesta en **explotación del modelo**

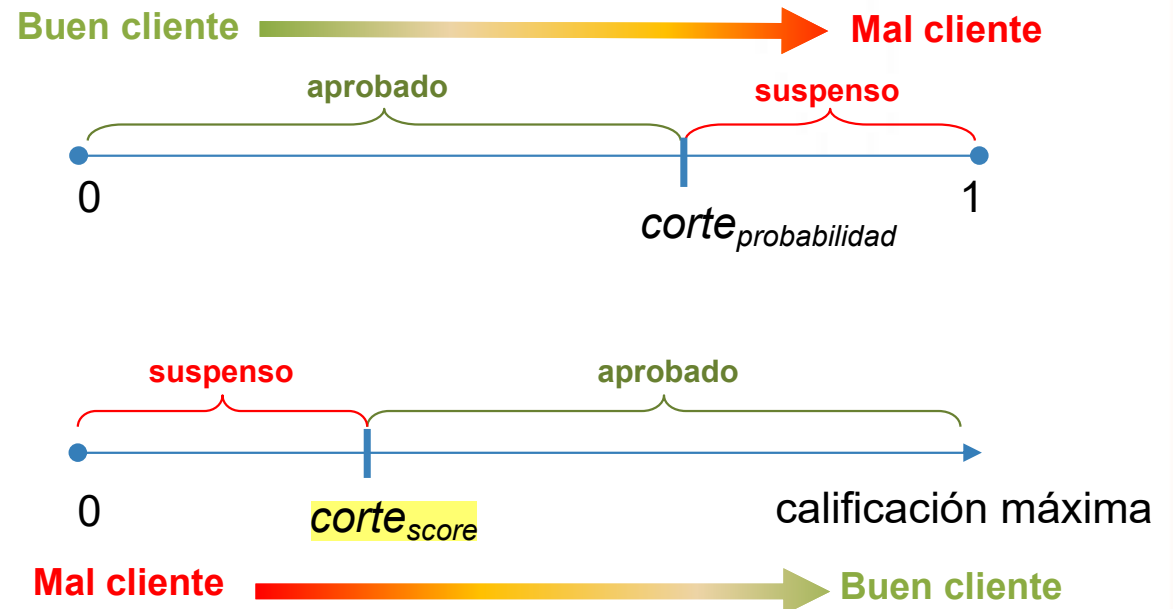
Probabilidades (o odds ratios) VS Puntos (Score)

El modelo estimado es un modelo de probabilidad y proporciona la probabilidad de impago, esto es, de ser mal pagador, o tal vez, en el caso de la regresión logística, el odd ratio



Probabilidad de impago

Score o Calificación crediticia



Score o puntuación baja = alto riesgo de impago
Score o puntuación alta = buen cliente bajo riesgo de impago

La probabilidad es tanto mayor cuanto más malo sea el cliente. Pero queremos una medida del riesgo, una puntuación que vaya al revés, que cuanto mejor sea la puntuación mejor sea el cliente. Queremos una medida de la calidad crediticia

$$\text{Score} = \text{Puntuación máxima} (1 - P(\text{impago}))$$

Ecuación para transformar $\ln(\text{odds})$ en Score

Se utiliza una transformación lineal arbitraria

$$\text{Score} = \text{Offset} - \text{Factor} * \ln(\text{odds})$$

¿Offset? ¿Factor?

Se establecen a partir de un punto arbitrario y una pendiente también arbitraria **por ejemplo:**

- (la pendiente) cada 20 puntos se doblan los odds ratio ($\text{pdo}=20$)

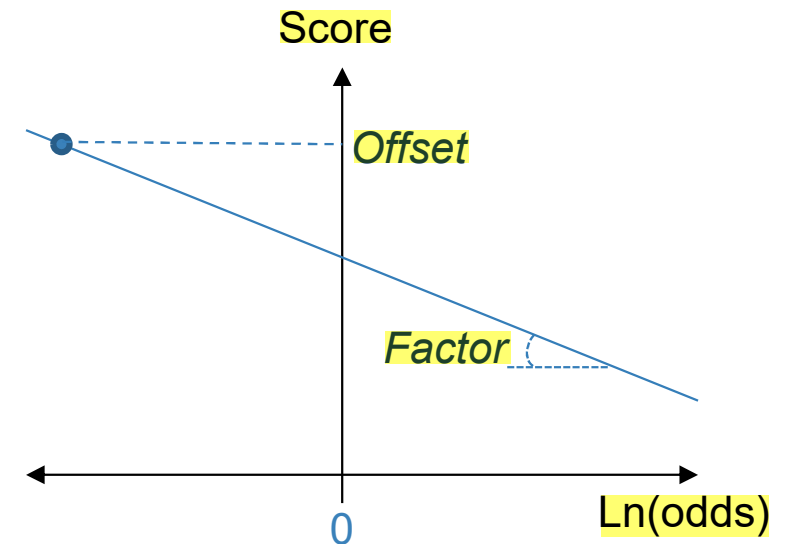
$$\text{Score} - 20 = \text{Offset} - \text{Factor} * \ln(2 * \text{odds})$$

$$\text{Score} - 20 = \text{Offset} - \text{Factor} * \ln(\text{odds}) - \text{Factor} * \ln(2)$$

$$20 = \text{Factor} * \ln(2) \longrightarrow \text{Factor} = 20 / \ln(2)$$

- (el punto) alguien que tenga un odd ratio de 1:50 tendrá 600 puntos

$$600 = \text{Offset} - \text{Factor} * \ln(1/50) \longrightarrow \text{Offset} = 600 + [20 / \ln(2)] * \ln(1/50)$$



En el caso de la **regresión logística** se puede transformar el modelo en forma de logaritmo del Odd Ratio para obtener un modelo lineal, donde el riesgo, medido por el logaritmo de los odds, es una combinación lineal de las variables explicativas

$$P(y=1|x_1, x_2, x_3, \dots, x_m) = \frac{1}{1 - e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m)}} \longrightarrow \text{Logit}(P) = \ln(P/(1-P)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m$$

Riesgo como suma de componentes

Esto permite expresar el score o puntuación de la calidad crediticia como una suma. Aunque ahora la transformación desde riesgo (medido por log(odds)) a score requiere establecer arbitrariamente dos parámetros

$$\text{Logit}(P) = \ln(P/(1-P)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m \longrightarrow \text{SCORE}$$

$$\text{Score} = \text{Offset} - \text{Factor} * \ln(\text{odds})$$

Variable	Atributo	Puntuación
Edad	Menor < 23	63
Edad	23-28	76
Edad	28-34	79
Edad	34-46	85
Edad	46-51	94
Edad	51- Mayor	105
Tipo Tarjeta	AMEX, VISA, Sin TRJ	80
Tipo Tarjeta	MasterCard	99
Salario	Menor <600	85
Salario	600- 1200	81
Salario	1200- 2200	93
Salario	2200 > Mayor	99
Estado Civil	Casado	85
Estado Civil	Resto	78

Ecuación para transformar ln(odds) en Score

$$\text{Logit}(P) = \ln(P/(1-P)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m \rightarrow \text{SCORE}$$

$$\text{Score} = \text{Offset} - \text{Factor} * \ln(\text{odds})$$

$$\text{Factor} = 20 / \ln(2)$$

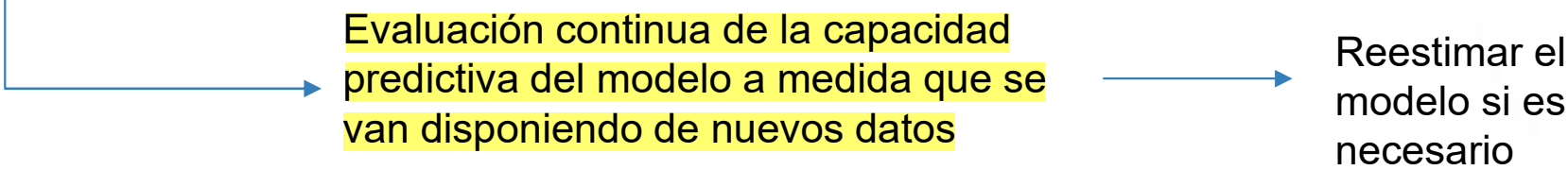
$$\text{Offset} = 600 + [20 / \ln(2)] * \ln(1/50)$$

- Cada 20 puntos se doblan los odds ratio (pdo=20), y (el punto) alguien que tenga un odd ratio de 50:1 tendrá 600 puntos

$$\text{Score} = \text{Offset} - \text{Factor} * \beta_0 - \text{Factor} * \beta_1 x_1 - \text{Factor} * \beta_2 x_2 + \dots - \text{Factor} * \beta_m x_m$$

WoEs se define $(\ln(B_i/B) - \ln(G_i/G))$ [coeficientes β positivos]

Una vez concluida la diagnosis y construida la **tarjeta de puntuación** y puesto en **explotación del modelo**

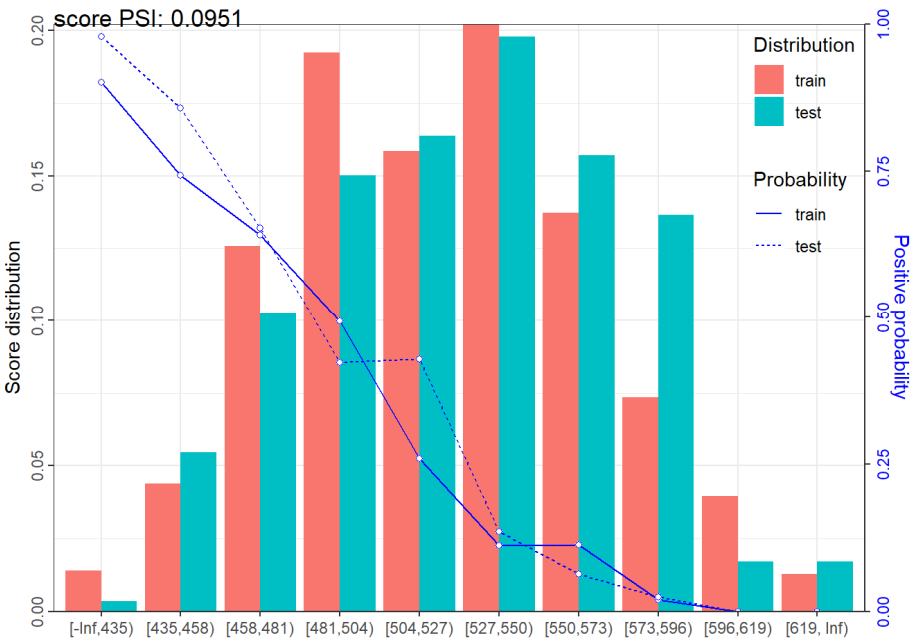


Estadístico **PSI** (Population Stability Index)

Es una medida de diferencia en la distribución de dos muestras, en nuestro caso entre la muestra utilizada para construir el modelo (entrenar y validar el modelo), y los nuevos datos que se vayan obteniendo con el transcurso del tiempo. Se aplica para detectar cuándo comienzan a verse diferencias entre las dos muestras (las puntuaciones de la muestra -train- y las puntuaciones obtenidas con los nuevos datos Cuando las distribuciones dejen de parecerse será el momento de revisar el modelo a tenor de los nuevos datos

$$PSI = \sum \left[\left(\frac{nuevos_i}{total\ nuevos} - \frac{train_i}{total\ train} \right) \cdot \ln \left(\frac{\frac{nuevos_i}{total\ nuevos}}{\frac{train_i}{total\ train}} \right) \right]$$

- **PSI <0.1:** No hay diferencias significativas entre las muestras de entrenamiento y los nuevos datos (resultado deseado, no se requiere más acciones)
- **PSI entre 0.1 y 0.25** Hay cambio menores, valdría la pena revisar el modelo
- **PSI >0.25** hay cambios importantes entre las dos muestras **HAY QUE CAMBIAR EL MODELO**



Ahora..... Vamos a ver alguna práctica

UNIVERSIDAD
COMPLUTENSE
DE MADRID

