UNIVERSIDAD COMPLUTENSE DE MADRID

ntic master
revolucionamos la comunicación

Módulo: Minería de datos y modelización

Guía práctica para la construcción de modelos de valoración de riesgo de Crédito (Scoring de Riesgos)

Autor: Lorenzo Escot

Profesor de la Facultad de Estudios Estadísticos

Guía práctica para la construcción de modelos de valoración de riesgo de Crédito (Scoring de Riesgos)

1. INTRODUCCION.

El **objetivo del Scoring o Puntuación de Riesgos** es **valorar el riesgo de impago** (probabilidad de impago o probabilidad de incumplimiento) de un cliente en caso de que se les conceda un préstamo (la técnica es muy parecida a la utilizada en los modelos de fraude, o de fuga de clientes o de tarificación de seguros). Esta calificación crediticia del deudor permitirá **poner nota a los posibles clientes**, de forma que podamos separar a los **malos clientes** (clientes con baja nota crediticia, *bajo score* o baja puntuación) de los **buenos clientes** (clientes con una buena nota crediticia, *alto score* o alta puntuación), de forma que sólo se concederán créditos a los clientes que aprueben esta calificación. Con estos modelos se intenta anticipar las situaciones de impago.

1.1 MODELOS DE PROBABILIDAD

Detrás de estos modelos de puntuación del riesgo de impago se encuentran los modelos de probabilidad. En estos modelos **la variable objetivo es la probabilidad de que ocurra un impago**, y se trata de estimar dicha probabilidad en función de una serie de variables o características observadas de los clientes.

$$P(impago) = F(\text{Características del individuo en el momento de solicitar el préstamo})$$

Para estimar estas probabilidades hay que seleccionar una muestra de clientes a los que ya se les ha concedido un crédito, algunos de ellos habrán pagado bien (*impago*=0), y otros habrán pagado mal (*impago*=1, habrán incurrido en impago).

La estimación del modelo consiste en entrenar con esa muestra el modelo para que por **semejanza** se le pueda asignar a un nuevo cliente una probabilidad de impago. Es decir, obtener un modelo que pronostique la probabilidad de que un cliente, en función de sus características, sea un mal pagador (mal cliente) en caso de que se le conceda un crédito.

Posteriormente, una vez obtenido el modelo, se aplicará a los nuevos clientes. A un nuevo cliente al que el modelo le asigne una probabilidad muy alta de ser mal pagador (de que haga impago) no deberíamos concederle el crédito. Por el contrario, a un nuevo cliente al que el modelo le asigne una probabilidad muy baja de hacer impago entonces sí que se podría conceder el crédito, ya que no se anticipa que ese cliente vaya a incurrir en impago.

1.2 SESGOS DE SELECCIÓN MUESTRAL

Los modelos de probabilidad de impago se entrenan con clientes a los que ya se les ha concedido un préstamo, y se sabe si han pagado bien(*impago*=0) o han pagado mal (*impago*=1). Estas estimaciones de la probabilidad de impago serán por tanto una medida relativa de impago, en relación con el conjunto de individuos analizados. Por eso en este tipo de modelos es fundamental analizar (y corregir) los **sesgos de selección muestral** que hacen que los modelos realicen una mala predicción del riesgo, es decir, que **infravaloren (o sobrevaloren) de manera sistemática la verdadera probabilidad de impago.**

El sesgo de selección muestral aparece cuando se selecciona una **muestra de la población para entrenar mi modelo predictivo que en realidad no es representativa de toda la población sobre la que quiero aplicar el modelo.**

El modelo será válido y funcionará bien sólo cuando la muestra de clientes que se esté utilizando para entrenar el modelo sea similar y representativa de todo el conjunto de individuos sobre el que después se aplicará el modelo de puntuación del crédito.

El modelo siempre funcionará bien cuando lo aplique sobre la submuestra de la población que sí representa a la muestra seleccionada. Sin embargo, cuando quiera aplicar el modelo al conjunto de toda la población se cometerán errores y no será válida la puntuación de riesgo si la muestra de entrenamiento es sistemáticamente diferente a la población objetivo, es decir, si los datos para entrenar el modelo están sesgados (no representan a toda la población).

El sesgo de selección muestral aparece sobre todo en los **modelos de riesgo de admisión**, esto es, en los modelos de valoración de riesgo de nuevos clientes. Nuevos clientes sobre los que no disponemos de más información que la que le solicitamos en el momento de realizar su puntuación crediticia. Ello es debido a que para entrenar a estos modelos se utiliza la muestra de clientes reales, individuos aceptados como clientes, individuos a los que alguien decidió en su momento conceder el crédito. Posteriormente algunos de estos clientes habrán pagado mal (impago=1) y otros bien (impago=0), y con esa información se entrena el modelo. El problema es que todos esos clientes son clientes que ya han pasado un filtro inicial (seguramente con algún modelo previo de valoración de riesgo), y son potencialmente buenos clientes.

Por el contrario, los nuevos clientes que acudan a la entidad para solicitar ser admitidos no son sólo los potencialmente buenos, también acudirán los potencialmente malos, que son los que precisamente deseamos detectar con el modelo de puntuación del riesgo. A estos últimos malos clientes se les estará sobrevalorando su calidad crediticia, porque se les está aplicando el modelo de probabilidad estimado con los buenos clientes potenciales

El problema de la selección muestral se encuentra, por tanto, en que, por un lado, la población sobre la que se desea aplicar el modelo, los nuevos clientes que solicitan un préstamo, estará compuesta por potenciales buenos y potenciales malos clientes. Mientras

que, por otro lado, el modelo ha sido entrenado sólo con buenos clientes (los clientes ya previamente aceptados como clientes). Esto es, en la estimación del modelo no se estarán utilizando a los clientes rechazados, los clientes que solicitaron un crédito, pero a los cuales se les rechazó el crédito por considerar que tenían una baja calidad crediticia, y que por tanto eran potencialmente malos clientes. Y no se pueden utilizar a los clientes rechazados para estimar el modelo porque como fueron rechazados, no se les concedió el crédito, y no se tiene, por tanto, información de su variable dependiente. No se observa la variable *impago* de los rechazados, sencillamente porque no se les concedió el crédito.

Las técnicas para solucionar este problema de selección muestral, como veremos más adelante consisten precisamente en inferir de alguna manera que hubieran hecho estos clientes rechazados en caso de que se les hubiera concedido el préstamo, lo que es conocido como **inferencia de rechazados**, para incorporarlos posteriormente a la muestra de aceptados y estimar el modelo de puntuación de riesgo sin sesgo de selección muestral.

1.3 Probabilidad de impago vs puntuación de riesgo

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 impagado, y 0 en caso contrario. Por lo tanto, los modelos de probabilidad que se analizarán querrán estimar la probabilidad P(*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 incurrido en impago (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 (aquéllos que han incumplido con sus obligaciones de pago) será un mal modelo para medir el riesgo de crédito.

La mayoría de los 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 alta puntuación o score de riesgo de crédito. Cuanta más puntuación (nota) mejor calidad del cliente. Y al contrario, un mal cliente (con alta probabilidad de impago) será aquel que tenga muy baja puntación o score (suspenderá el examen de evaluación de su calidad crediticia)

2. FASES PARA CONSTRUCCION DE UN MODELO DE ADMISION

Las fases para la construcción de un modelo de valoración son las siguientes:

1) **Construcción y depuración inicial de la base de datos**
2) **Definición de la variable impago ¿Cuándo se hace impago?**
3) **Selección de Variables Explicativas ¿Qué Variables debería seleccionar como variables explicativas?**
4) **Tramificación de las variables cuantitativas y agrupación de categorías**
5) **Transformación WoE de las variables**
6) **Estimación y Diagnosis del modelo de probabilidad**
7) **Inferencia de denegados y estimación del modelo final**
8) **Transformación de la probabilidad en puntuación: Tarjetas de Puntuación**
9) **Puesta en producción del modelo**

Vamos a continuación a comentar los elementos principales de cada una de esas fases

2.1 Construcción y depuración inicial de la base de datos

En muchas de las aplicaciones reales, las principales dificultades se encuentran precisamente en la obtención de los datos que se van a utilizar para estimar el modelo, que en muchos casos se encuentra dispersa en diferentes departamentos y bases de datos.

Una vez construcción esa base de datos con los campos o variables que potencialmente podrían estar explicando la probabilidad de impago, habrá que hacer una depuración inicial de la misma:

- ¿Tiene la variable un excesivo número de registros perdidos (en blanco)?
- ¿Tiene alguna de las variables un excesivo número de datos anómalos?
- ¿Hay alguna variable para la que sus datos no varíen lo suficiente (toman siempre el mismo valor para todos los registros o individuos)?

Con este tipo de análisis inicial despreciaremos algunas de las posibles variables de la base de datos, ya que no proporcionarán información relevante para construir el modelo. Y en cualquier caso, habrá que tomar una serie de decisiones para las variables que finalmente se seleccionen como potencialmente integrantes del modelo de probabilidad:

- ¿Hay que intervenir los datos anómalos? Por lo general sí, sobre todo si no tienen sentido (por ejemplo créditos concedidos a menores de edad, o importes de crédito negativos, …)
- ¿Hay que imputar los datos perdidos? Por lo general, no. Es casi preferible dejar los datos perdidos como una categoría más y analizar si esa categoría resulta significativa en los modelos de probabilidad, esto es, que el hecho de que un cliente no proporcione información sobre alguna variable sí está relacionado con la probabilidad de que ese cliente vaya a incurrir en impago.
- ¿Hay que transformar las variables? Por lo general tendremos variables categóricas y variables numéricas. Con las categóricas trabajaremos inicialmente sin más transformaciones, pero algunas variables numéricas, sobre todo las que miden salarios, rentas, importes, precios, etc.… suelen aceptar la transformación logarítmica (se distribuyen como una log-normal). Pueden utilizarse diferentes pruebas estadísticas para saber cuándo es necesario transformar estas variables para que se acerque más a la normalidad

2.2 Definición de la variable impago ¿cuándo se hace impago?

Una de las variables que debemos definir es la propia variable objetivo, la variable *impago*. Esta será una variable dicotómica, *impago*=1 cuando el cliente ha hecho impago, *impago*=0 cuando no ha hecho impago.

En este punto es necesario establecer:

- ¿Cuánto tiempo se espera desde la fecha en la que se debería haber hecho el pago hasta que se declara esa operación como de impagada? ¿un mes, dos meses, tres meses?
- ¿Durante cuánto tiempo sigo a un crédito para calificarlo como de impagado?, ¿sigo al cliente durante toda la vida del crédito, hasta que llega a vencimiento? ¿cuál es la ventana de observación?

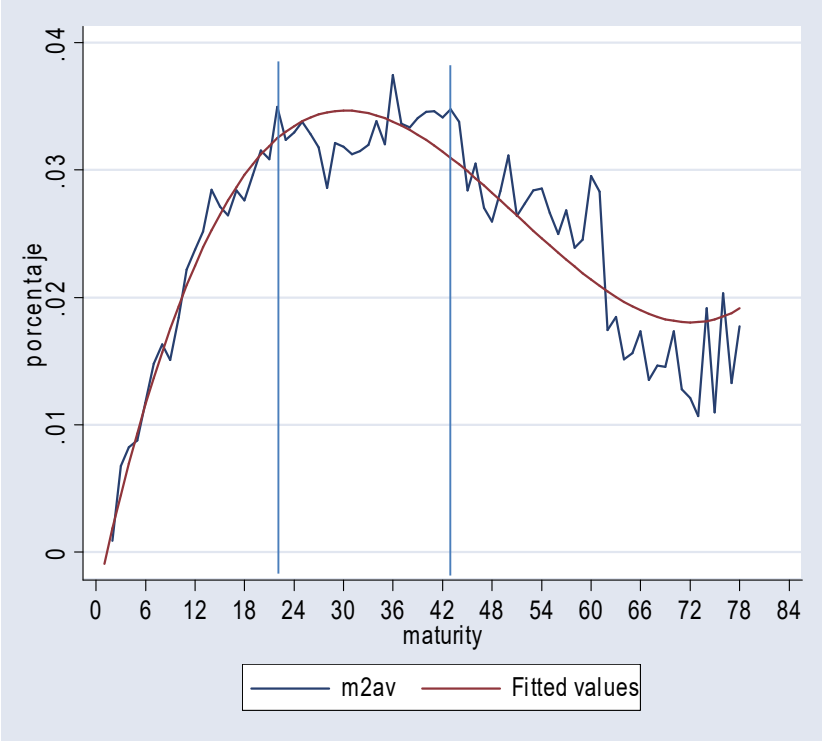
Ambos puntos deben ser definidas porque el modelo estimará la probabilidad de impago en un día, o un mes, o tres meses o cuatro meses, y durante los 6 primeros meses, durante los dos primeros años o durante toda la vida del préstamo, dependiendo de cómo se haya definido esta variable objetivo.

Figura 1. Periodos de Mora para la definición de la variable impago

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

A modo ilustrativo, el Banco de España define una operación como impagada a partir de los 3 meses de mora. En cuanto a la ventana de observación suelen utilizarse los primeros meses de la vida del préstamo, porque es cuando se alcanzan los máximos porcentajes de impagos (figura 2), y porque además para construir el modelo se va a utilizar la información disponible al principio del préstamo, en el momento en el que el cliente solicita el crédito, para predecir si el cliente va a ser buen o mal pagador en el futuro. La ventana de observación determina exactamente cuánto de futuro. Previsiblemente cuanto más lejano sea ese futuro menor relación habrá entre las condiciones iniciales del cliente y su calidad crediticia futura, por eso no suelen utilizarse ventanas demasiado alejadas del momento inicial del préstamo

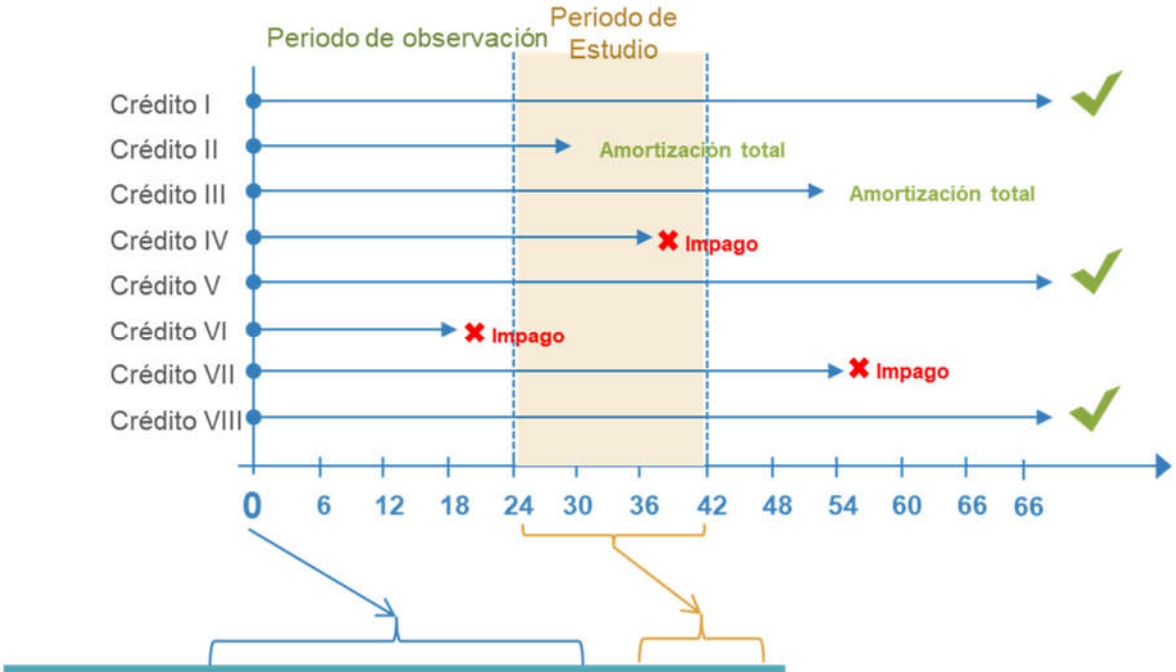
Figura 2. Porcentaje de operaciones impagadas en función del número de meses desde el inicio del crédito (maduración)



Con la información de esta variable objetivo impago, y la información sobre las características del cliente disponible en el momento inicial (porque estamos en un modelo

de admisión, y la única información disponible sobre el cliente es la que la disponible en el momento en el que se solicita el crédito) procederemos a construir el modelo de valoración del riesgo

Figura 3. Ventana de observación para la construcción de la variable impago



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

2.3 Selección de Variables Explicativas ¿qué variables debería seleccionar como variables explicativas?

Una vez que se tiene definida la variable objetivo impago, y una vez que tenemos la base con todas las potenciales variables, comenzamos a construir el modelo propiamente dicho.

El primer paso es separar el conjunto de datos en dos submuestras, la **muestra de entrenamiento y la muestra de validación** para realizar una diagnosis de nuestro modelo. Típicamente un 70%-75% de los datos se dejan para el entrenamiento y un 30%-25% para la validación del mismo.

A partir de la muestra de entrenamiento debemos seleccionar las variables, del conjunto de todas las posibles variables explicativas, cuáles de ellas se incluirán en el modelo. No deben ser muchas, las suficientes para construir un modelo que permita separar bien, discriminar bien a los buenos y a los malos clientes. Téngase en cuenta de estamos en modelos de admisión y que por tanto debemos preguntar por ellas a cada nuevo cliente que viene a solicitarnos un crédito, y por tanto tenemos que elegir muy bien que le preguntamos al cliente.

Para realizar este tipo de selección o cribado inicial de variables existen dos alternativas (no excluyentes) que podemos utilizar para responder a la cuestión de ¿qué variables de las posibles están estadísticamente relacionada con la variable objetivo y me servirían para separar a los buenos d ellos malos clientes?: una basada en criterios de asociación y otra basada en criterios de asociación.

Por un lado, se pueden utilizar técnicas de asociación bivariante. Es bivariante porque se va analizando, una a una, la asociación entre cada variable potencialmente explicativa y la variable objetivo *impago*. Téngase en cuenta que como la variable *impago* es dicotómica, la técnica para medir asociación con ésta dependerá del tipo de variable explicativa que se tenga en cada caso. Por ejemplo, si la variable explicativa es también categórica (sexo, nacionalidad, nivel de estudios, estado civil, etc.…) se podrían utilizar test de independencia tipo Chi-cuadrado. Si por el contrario las variables explicativas son cuantitativas (renta, salario, edad, importe total del préstamo solicitado, etc.) se podría hacer algún contraste de diferencia de medias.

Una vez contrastada la hipótesis de ausencia de asociación entre cada posible variable explicativa y la variable objetivo impago, seleccionaremos inicialmente sólo aquellas que tengan una asociación significativa con la variable objetivo. Estos contrastes de asociación bivariante van perdiendo potencia a medida que aumenta el tamaño muestral (tienden a detectar asociación significativa de todas las variables con la variable *impago*), por eso, en los análisis de riesgos se suele utilizar un segundo criterio para la selección de variables.

Ese segundo criterio de selección de variables se basa no en medidas de asociación, sino en medidas de concentración. Se buscan variables explicativas que sean categóricas o cualitativas, con diferentes niveles o categorías (ejemplo, nivel de estudios: sin estudios, secundaria obligatoria, estudios superiores) que concentren a muchos buenos clientes o a muchos malos clientes en alguno de sus niveles o categorías. De manera que, si un cliente cae en esos niveles o categorías que concentra a muchos malos pagadores, pueda asignarle una alta probabilidad de que vaya a ser también mal pagador.

Imagínese por ejemplo que se está analizando el estado civil que definimos con 6 diferentes niveles o categorías: soltero, conviviendo en pareja, casado, separado, divorciado, y viudo. Si el porcentaje de impagos en cada uno de esos 6 niveles fuese el mismo, entonces no habría concentración de impagados en ninguna categoría, y por tanto, conocer el estado civil de un cliente no ayudaría, no aportaría información, para saber si va a ser buen o mal cliente. Sin embargo, si detectamos que los malos clientes están concentrados todos entre los divorciados (es sólo un ejemplo), conocer el estado civil de un individuo sí que aportaría mucha información para pronosticar si será mal o buen cliente. Por semejanza, en caso de que el individuo estuviese divorciado debería hacernos sospechar que será mal pagador. Y al contrario, si el solicitante no está divorciado, podríamos clasificarlo como de buen cliente, porque no existen impagos entre los no divorciados.

Este es un caso extremo porque en realidad es difícil encontrar una variable categórica que concentre a todos los malos pagadores. Así que normalmente se utilizan diferentes estadísticos que ayudan a medir esa concentración: El Information Value y el índice de Gini

- **Valor de la información** (***Information Value o IV***). Es una medida de la información que proporciona una variable categórica para separar a los buenos de los malos clientes que se basa en la concentración relativa de buenos y malos clientes en cada una de sus categorías. Esa concentración relativa de buenos y malos clientes en cada nivel o categoría se mide por otro estadístico denominado *WoE* (del inglés Weight of Evidence) de la categoría o nivel. Así para una variable categórica que tenga 6 categorías existirán 6 medidas de WOE:

$$WoE_i = ln\left(\frac{\frac{Malos_i}{Total\ Malos}}{\frac{Buenos_i}{Total\ Buenos}}\right) \qquad i = 1,2\ldots,6$$

  Fíjese que cuanto mayor es el *WoE* de una categoría mayor concentración relativa de Malos clientes habrá en esa categoría (relativa porque se mide en relación con la concentración de buenos clientes en esa categoría). Por ejemplo si en una categoría se concentrase el 5% del total de malos clientes y el 5% del total de buenos clientes, el *WoE* de esa categoría sería cero (ln(1)=0). Esa categoría me da poca información. Utilizar criterios de selección de variables explicativas basados en concentración consiste en buscar, por tanto, variables categóricas o cualitativas que tengan alguna de sus categorías con *WoE* muy altos (concentren relativamente a muchos malos clientes) o con *WoE* muy bajos (concentren relativamente a muchos buenos clientes).

Para resumir la información de todos los *WoE* de una variable categórica, se calcula el estadístico *Information Value* (*IV*):

$$IV = \sum_{i=1} \left(\frac{Malos_i}{Total\ Malos} - \frac{Buenos_i}{Total\ Buenos} \right) WoE_i$$

Nótese que este *IV* es único para una variable categórica, y no es más que una media ponderada de los *WoE* de sus diferentes niveles o categorías, donde las ponderaciones están elegidas para que no se anulen sus WoE positivos con sus WoE negativos. Cuanto mayor sean los WoE en valor absoluto (mayor concentración relativa de buenos o malos clientes) de las categorías de una variable categórica, mayor será el *IV* de dicha variable, y más información proporcionará para separar a los buenos de los malos clientes (véase ejemplo de cálculo en la figura 4).

Figura 4. Ejemplo de cálculo del WoE y de Valor de Información

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**

La selección de variables explicativas basadas en el IV exige que ese IV sea, por tanto, elevado, ¿cuánto de elevado? Estos criterios de concentración no se basan en contraste de hipótesis tradicionales, ni utilizan el concepto de p-valores de la inferencia estadística tradicional, pero tampoco siguen ninguna regla fija a la hora de decidir cuándo seleccionar (o no seleccionas) una posible variable explicativa para ser incluida en el modelo predictivo. Suele utilizare el umbral (flexible) de 0.02: s**i el IV de una variable es menor que 0.02, la variable se considera no predictiva**, y no debería utilizarse en el modelo; si el IV es mayor que 0.3 es muy predictiva (Siddiqi, 2006).

- ***Índice de Gini***. Otro estadístico de concentración también utilizado como criterio de selección es el bien conocido índice de Gini que toma valores entre 0 (cuando todos los niveles o categorías presentan el mismo porcentaje relativo de malos clientes) y 1 (cuando todos los malos se encuentran concentrados en una única categoría o nivel de la variable. Para construir el índice de Gini primero deben ordenarse las m categorías en orden descendente por la proporción de malos en cada una de ellas. Esto es importante porque la fórmula del índice de Gini requiere calcular frecuencias acumuladas:

$$Gini = \left(1 - \frac{2 \times \sum_{i=2}\left(Malos_i \times \sum_{j=1}^{i-1} Buenos_j\right) + \sum_{i=1}(Malos_i \times Buenos_i)}{Total\ Malos\ \times Total\ Buenos}\right) \times 100$$

Cuanto mayor sea el índice de Gini mayor concentración de malos clientes en alguna de las categorías de la variable, y por tanto más información proporciona esa variable para separar a los buenos de los malos clientes. De nuevo no existe una regla fija, pero suele aceptarse como criterio de selección que el índice de Gini sea superior a 0.15 (Siddiqi, 2006).

Terminamos recordando que en realidad no existe una regla fija para seleccionar una variable como explicativa utilizando criterios de concentración (IV o Gini). Es por este motivo que en muchos casos se analizan diferentes criterios conjuntamente para tomar la decisión final de qué variables incluir en el modelo, no sólo criterios de concentración, también de asociación o incluso, y no menos importantes, criterios de negocio (el conocimiento del negocio, de qué variables funcionan y qué variables no funcionan es también muy importante) que incluyan la no repetición de variables que proporcionen esencialmente la misma información (por ejemplo, el número de miembros del hogar, los ingresos totales del hogar, y los ingresos per cápita del hogar). De forma adicional, una vez realizada esta selección inicial, también pueden utilizarse algoritmos de estimación que incluyan procesos de selección de variables por pasos (hacia adelante, hacia atrás o combinados), que por lo general proporcionarán las mismas variables de selección, aunque sí que serán capaces de detectar y filtrar variable que se encuentren altamente correlacionadas entre sí.

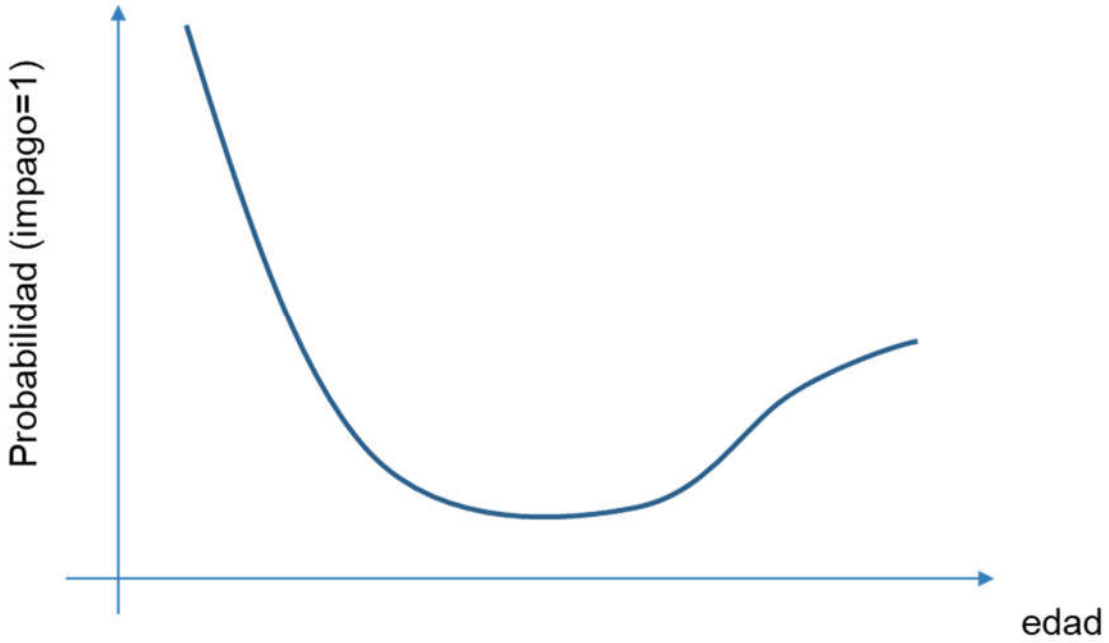
2.4 Tramificación de las variables cuantitativas y agrupación de categorías

En gestión de riesgos, 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, ya que si la relación entre la variable impago y la variable continua es no lineal, los modelos lineales de probabilidad no van a conseguir estimar un buen modelo (el ajuste será muy malo). Si la relación entre una variable cuantitativa y la probabilidad de impago es no lineal, entonces hay dos opciones. O dejar de usar modelos lineales para estimar la probabilidad (porque no son capaces de captar las no linealidades) o bien categorizar, o discretizar o tramificar la variable categórica de forma que podamos recoger esa no linealidad, midiendo el efecto de cada categoría o nivel sobre la probabilidad por separado

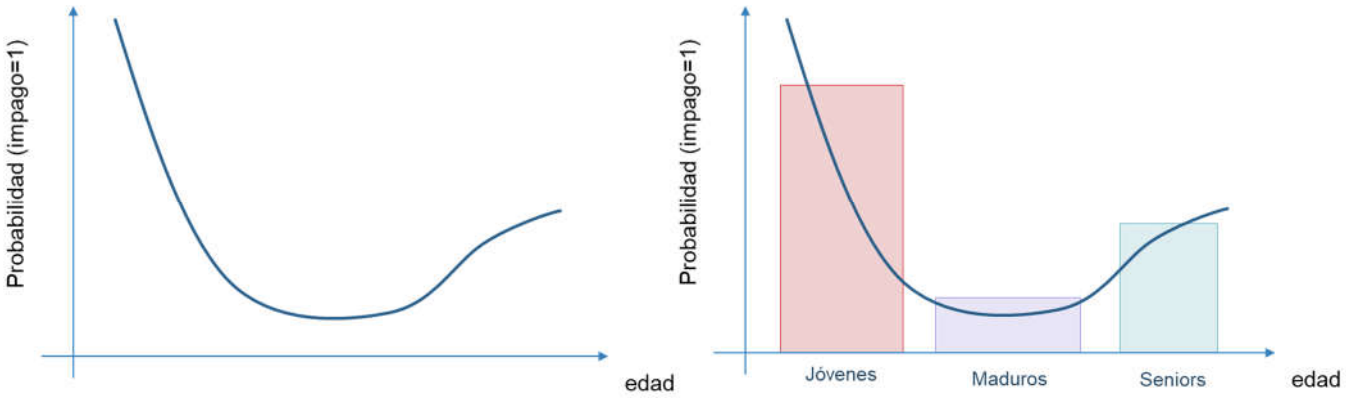
Por ejemplo, suponga que la relación entre la edad y la probabilidad de impago es no lineal: los más jóvenes son más propensos al impago; a medida que se van cumpliendo años nos volvemos mejor pagadores y se reduce la probabilidad de impago; y ya a más avanzada edad vuelve a subir ligeramente la probabilidad de impago (Figura5)

Figura 5. Relación no-lineal entre Probabilidad de impago y edad



Entonces no podemos ajustar esta relación entre **edad y la probabilidad de impago mediante un modelo lineal** (figura 6A). Para captar la no linealidad mediante un modelo lineal habría que convertir 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 (figura 6B)

Figura 6 Ajuste de la relación no lineal con modelos lineales con uso de variables cuantitativas vs categóricas



Caso A: edad como variable

Caso B: edad como variable

Existen diferentes técnicas para hacer esta tramificación de variables cuantitativas 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.

En los modelos de riesgo se busca además simplificar en lo posible las variables categóricas que se incluirán en el modelo. No solo se quiere que el número de variables sea lo menor posible, sino además que la definición de esas variables sea también lo más simplificada posible, es decir, se busca, para mejorar la interpretabilidad, que las variables categóricas que se vayan a incluir en el modelo tengan pocos niveles o categorías, los suficientes para que la variable sea informativa, pocos niveles, por tanto, que concentre a muchos buenos o malos clientes en alguna de esas categorías.

Por tanto, en los modelos de riesgo, no sólo se va a tramificar o discretizar a todas las variables continuas, también se va a agrupar todas aquellas categorías de todas las variables categóricas (de las que inicialmente ya eran categóricas y de las que eran inicialmente continuas pero hemos categorizado) que proporcionen la misma información, simplificando el número de niveles o categorías de una variable siempre que el IV (o el Gini) total de la variable tramificada y agrupada vaya aumentando. Este proceso de **recodificar o reagrupar categorías de las variables categóricas** permite reducir el número de categorías que tendrán las variables categóricas y por tanto, simplificar el modelo.

2.5 Transformación WoE de las variables

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 (*IV* y *Gini*) y en su caso reagrupando de tramos o niveles de forma que la cantidad de información que proporcionen dichas categorías sea la máxima para poder discriminar entre buenos y malos clientes. De hecho, esas medidas de concentración se han utilizado para seleccionar las variables que finalmente se introducirán en el modelo (por ejemplo, IV>0,02 o $Gini$>0.15)

Una vez seleccionadas las variables que se introducirán en el modelo de probabilidad, la metodología de riesgos suele realizar una transformación adicional a las variables seleccionadas la **transformación WoE**, que consiste en **convertir las variables categóricas en variables continuas sustituyendo cada categoría nominal *i* de la variable original por su valor numérico WoE_i**

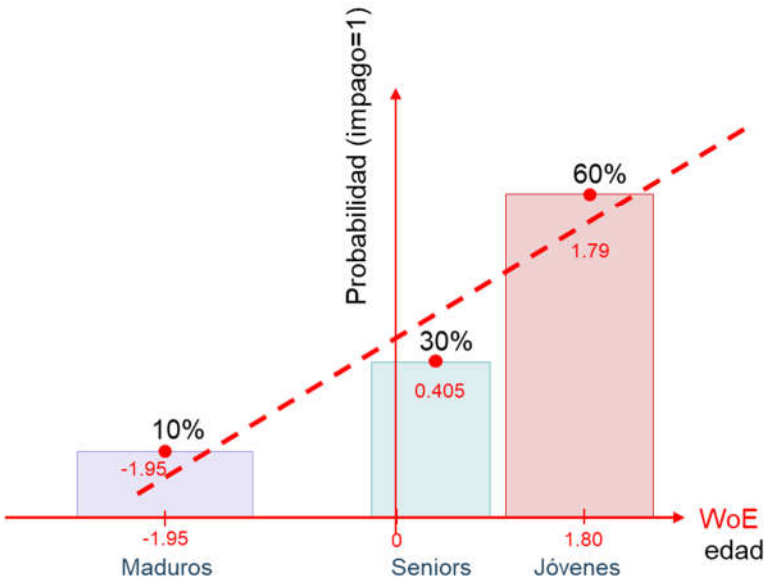
El principal motivo para realizar esta transformación WoE es de nuevo el de simplificar el número de variables que se introducirán en el modelo. La forma tradicional de incluir

una variable categórica con *m* categorías o niveles diferentes en un modelo de regresión es creando tantas variables binarias dicotómicas como niveles existan e introducir *m*-1 de dichas variables dicotómicas como variables explicativas del modelo. Eso supone multiplicar el número de variables que se introducirán en el modelo (se tendrán *m*-1 nuevas variables por cada variable categórica con *m* niveles o categorías), y por tanto el tamaño de la matriz de variables explicativas. Si en lugar de trabajar con variables categóricas, las transformamos para que sean continua, no sería necesaria la construcción de esas variables dicotómicas ni multiplicar el número de variables que se introducirán en el modelo. Se mantendrá el número de variables categóricas, con una única variable continua por cada variable categórica (aunque en realidad cada una de ellas solo tomará *m* diferentes valores, uno para cada una de las *m* categorías de la variable original).

Además, el uso de los **WOE permite linealizar todas las relaciones no lineales igual que lo haría el uso de diferentes variables dicotómicas**, 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, ya que dicho valor WoE vendrá determinado precisamente por el porcentaje relativo de Malos y Clientes, que es proporcional a la probabilidad de impago (en realidad al Odd-Ratio). 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 (figura 7).

Figura 7 Transformación WoE de las variables categóricas

	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

Figura 8 Resumen del procedimiento para la selección de variables utilizando medidas de concentración

Podemos resumir del procedimiento para la selección de variables utilizando medidas de concentración (Valor de información – IV - o índice de GINI) en los siguientes pasos:

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)

2.6 Estimación y diagnosis del modelo de probabilidad

Una vez que se tiene definida la variable objetivo impago, y una vez que tenemos la base con todas las variables explicativas que queremos introducir en el modelo (seleccionadas según algún criterio de concentración), y una vez que las hemos transformado en variables explicativas continuas WoE (Figura 8), procedemos a estimar el modelo de probabilidad propiamente dicho.

$P(impago) = F$ (Características del individuo en el momento de solicitar el préstamo)

Que escribiremos por comodidad, y de forma genérica como:

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

Donde y representa en nuestro caso el evento de impago, y por tanto cuando $y=1$ el cliente cometerá impago (mal cliente), mientras que cuando $y=0$ el cliente no hace impago (buen cliente) durante la ventana de observación; y donde se supone que existen k variables explicativas que ya han sido depuradas y transformadas debidamente y que están medidas de forma continua con la transformación WoE tal y como se ha expuesto en los apartados anteriores

Existen diferentes alternativas para la estimación de un modelo de probabilidad: desde el **modelo de lineal de probabilidad**, donde se hace depender la probabilidad de impago directamente de la combinación lineal de las variables explicativas:

$$P\,(y=1\mid \boldsymbol{x}^{WoE}) = \beta_0 + \beta_1\, x_1^{WoE} + \beta_2 x_2^{WoE} + \ \ldots + \beta_k\, x_k^{WoE}$$

o los **modelos de probabilidad lineales generalizados**, donde se utiliza una transformación no lineal de un índice lineal (una combinación lineal de las variables explicativas).

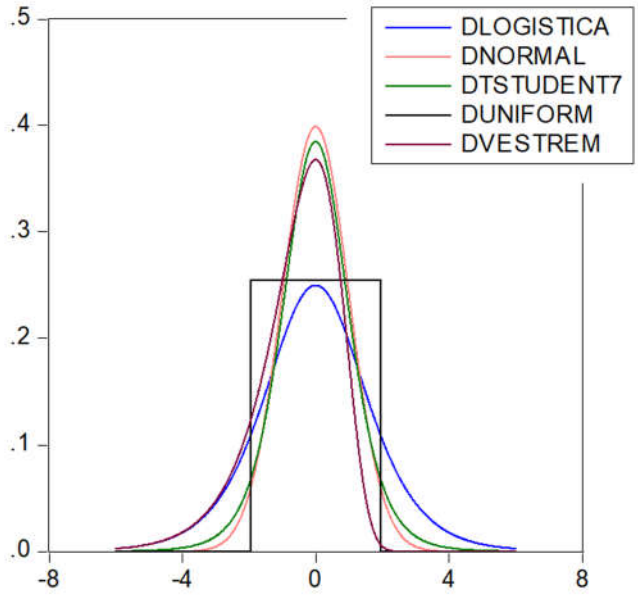
$$P\,(y=1\mid \boldsymbol{x}^{WoE}) = G(\,\beta_0 + \beta_1\, x_1^{WoE} + \beta_2 x_2^{WoE} + \ \ldots + \beta_k\, x_k^{WoE}\,)$$

Esta transformación no-lineal $G(\ )$ debe garantizar que la probabilidad estimada esté siempre entre 0 y 1, y por ello se utilizan funciones de distribución de probabilidad, y dependiendo del tipo de función de distribución los modelos recibirán una u otra denominación. Cuando se utiliza, por ejemplo, la Normal tenemos los *modelos probit*; y cuando se utiliza la Función de Vvalor extremo los modelos Gompit. Uno de los modelos más ampliamente utilizados en la gestión de riesgos la que utiliza como transformación $G(\cdot)$ la función de distribución logística, y esconocido como **modelo de probabilidad de regresión logística** (véase Wooldridge 2016

$$P\,(y=1\mid \boldsymbol{x}^{WoE}) = \frac{1}{1-e^{-(\boldsymbol{\beta}'\boldsymbol{x}^{WoE})}} =$$

$$= \frac{1}{1-e^{-(\beta_0+\beta_1\, x_1^{WoE}+\beta_2 x_2^{WoE}+\ \ldots+\beta_k\, x_k^{WoE})}}$$

Figura 9 Diferencia entre las diferentes funciones de distribución de probabilidad



La diferencia entre utilizar una función de distribución $G(\ )$ u otra se encuentra en el comportamiento de los valores extremos (en el individuo medio no existen tantas diferencias), siendo más anchas (más probables) los valores extremos en la logística que en la normal (Figura 9). Teniendo en cuenta que los porcentajes de impagados de las entidades suelen ser reducidos puede considerarse como algo extraordinario, por lo que suele preferirse la logística a la normal, por tener colas más anchas, pero también, como veremos más adelante porque la regresión logística es la única que permite linealizarse en los logaritmos de los *Odd-Ratios*, lo que permitirá construir tarjetas de puntuación con una fácil interpretación y cálculo. Pero esto se verá más adelante

En realidad, cualquier otro modelo predictivo de los que ya conocéis de podría utilizarse para ajustar y predecir las probabilidades de impago, desde los modelos de clasificación basados en árboles, random forest, SVM, algoritmos de Gradient boosting, los modelos de redes neuronales, o cualquier otro dentro del conocido como *maching learning y deep learning*.

La elección de uno u otro modelo debería fundamentarse en su capacidad predictiva (aunque como también hemos comentado anteriormente la interpretabilidad de los modelos puede ser un factor determinante, aspecto esto sobre el que volveremos más adelante). Por eso resulta fundamental realizar un diagnóstico apropiado de los modelos estimados, validando el comportamiento de los diferentes modelos fuera de su muestra para evitar el sobreajuste. En ese sentido recordemos que hemos reservado una muestra aleatoria de nuestros datos para realizar esta validación. Cualquier otro método, como los de validación cruzada permitirán comparar diferentes modelos predictivos para seleccionar el que finalmente utilizaremos para predecir la probabilidad de impago de nuestros futuros clientes.

No nos vamos a detener aquí en los procedimientos de estimación y ajuste de estos modelos. blemas (véase por ejemplo Aggarwal, 2015; Han et al. 2011; Hastie y Friedman, 2009; James et al. 2013; Kubat, 2017). Recordemos, simplemente alguna de las dificultades a las que se enfrentan estos modelos a la hora de realizar su diagnosis. En un modelo de regresión lineal, la bondad global de ajuste del modelo suele hacerse midiendo los errores de predicción, definidos como la diferencia entre el valor real y el valor ajustado. En los modelos de probabilidad no pueden utilizarse este tipo de técnicas porque no existen los errores como tales. Recordemos que la variable objetivo es la probabilidad de impago. Y la predicción que proporcionan estos modelos es sobre la probabilidad de impago. Sin embargo, la probabilidad de impago de un cliente no es una variable que pueda observarse, y por eso no puede construirse una variable de residuos como tal. No se observa la probabilidad a priori que tiene un cliente de ser mal pagador, pero sí se observa la realización a posteriori de este evento. Es decir, si se conoce si a posteriori un cliente fue buen o mal pagador ($y=0$ o $y=1$, respectivamente). Por eso para poder realizar una diagnosis de estos modelos es necesario convertir las probabilidades estimadas por el modelo predictivo estimado en pronósticos sobre nuestros clientes, ¿un cliente para el que se estime una probabilidad de impago de 0.20 pronosticaremos que será un buen o un mal pagador?

$$\begin{cases} cuando & \hat{P}_i \geq corte & pronostico\ que\ \hat{y}_i = 1 \\ cuando & \hat{P}_i < corte & pronostico\ que\ \hat{y}_i = 0 \end{cases}$$

Una vez realizados estos pronósticos podremos computar el porcentaje de nuestros pronósticos correctos e incorrectos. Estas medidas son las que se utilizan para hacer diagnosis de la bondad de ajuste del modelo. Hay que tener en cuenta por tanto que en el ajuste de estos modelos dependerá crucialmente del *corte* seleccionado para realizar los pronósticos. Existen diferentes criterios para elegir ese punto de corte (desde el punto para el que se alcanza el máximo valor del estadístico *K-S* o diferencia entre el número acumulado de malos y buenos clientes; hasta el máximo del *F-score*, o medida de ajuste que considera tanto la *precisión* como la *exhaustividad*), pero todos rondarán en torno a la frecuencia observada de eventos *y=1* en la muestra de entrenamiento. Nótese que por tanto la elección de un punto de corte = 0.5 sólo resultaría adecuado cuando en la muestra de entrenamiento exista un 50% de casos con *y=1* y otro 50% con *y=0* .

Una mala elección de ese punto de corte puede hacer que nuestro modelo, aunque esté bien ejecutado en todas sus fases de elección de variables y de estimación del modelo, resulte en un mal ajuste (medido por el porcentaje de pronósticos incorrectos). Así que es habitual utilizar otros estadísticos para la diagnosis de este tipo de modelos que no dependen crucialmente de la elección de un punto de corte. Por ejemplo, el área por debajo de la *Curva ROC*, curva que se calcula a partir de los pronósticos realizados no para uno, sino para muchos puntos de *corte* (por lo que no depende crucialmente de la decisión sobre un único punto de corte).

2.7 Inferencia de denegados y estimación del modelo final

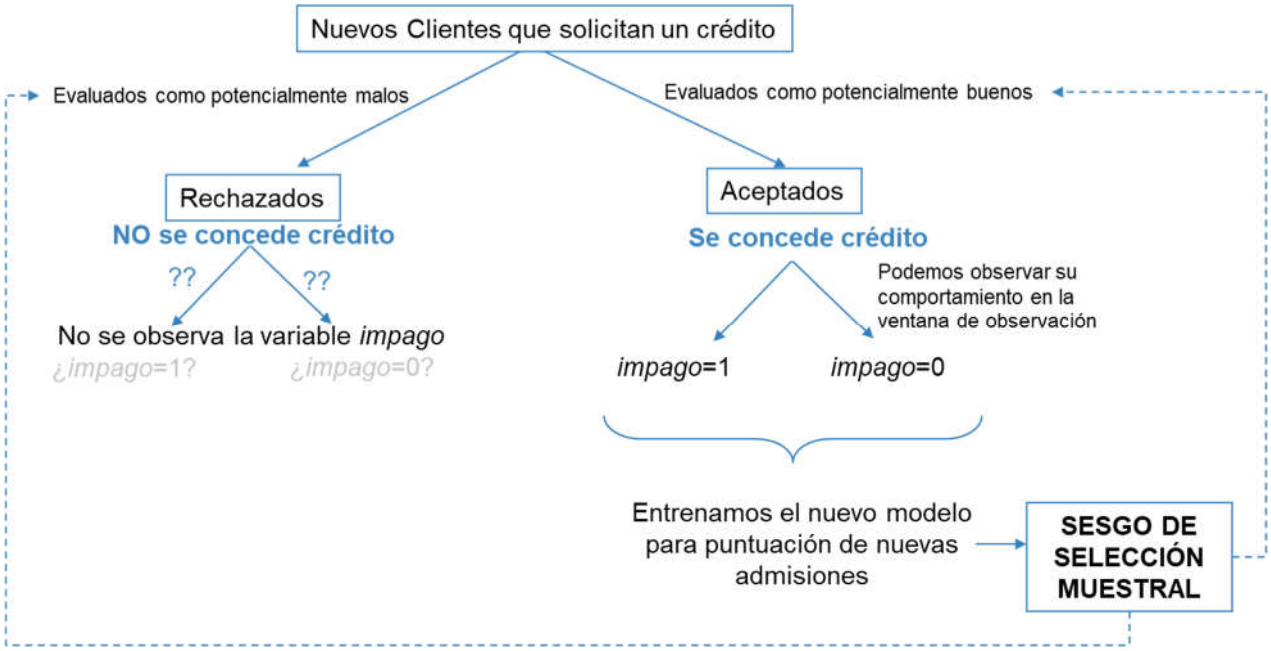
Ya hemos comentado previamente que uno de los problemas a los que se enfrenta el análisis de riesgos es el del **sesgo de selección muestral**, esto es, que la muestra de datos que utilicemos para entrenar nuestro modelo esté sesgada y no represente a toda la población sobre la que se querrá aplicar posteriormente el modelo predictivo de impago. En este sentido, una etapa fundamental en la diagnosis del modelo es el análisis de la posible presencia de este tipo de sesgos de selección muestral.

Este problema de selección muestral aparece sobre todo en los modelos de **scoring de admisión** en los que se tiene que evaluar solicitudes de nuevos clientes para su admisión, ya que, aunque sí es posible disponer de la información solicitada a todos los clientes que solicitaron un crédito (la requerida en su momento para evaluar su solicitud) que actuarán como posibles variables explicativas, en realidad sólo se dispone de información sobre la variable objetivo *impago* de los clientes que fueron aceptados, a los que se les concedió el crédito. Los clientes a los que se les rechazó la solicitud (probablemente porque se estimó en su día que tenían un algo riesgo de impago), los llamados clientes rechazados, no podrán incorporarse en la estimación del modelo porque no se observa para ellos la variable *impago*. No sabemos si estos clientes rechazados han incurrido en impago durante la ventana de observación, es decir, si son buenos o malos clientes, porque no les hemos concedido el crédito. Aunque hay que sospechar que si fueron rechazados es

porque alguien, utilizando algún modelo de valoración de riesgo previo, valoró que eran potencialmente malos clientes (figura 10).

El problema de este sesgo de selección es que cuando estimamos un modelo de puntuación de riesgo solo con los clientes aceptados (para los que sí se conoce si hicieron impago o no durante la ventana de observación), al aplicarlo sobre todas las nuevas solicitudes (que incluirán tanto a potenciales aceptados como a potenciales rechazados) estaremos asimilando que todas esas nuevas solicitudes son como las de nuestros clientes aceptados con los que hemos entrenado el modelo, que inicialmente alguien valoró como potencialmente buenos clientes (y por eso se les concedió el préstamo). Es decir, estamos sobrevalorando la calidad crediticia, a priori, de todos nuestros nuevos clientes considerándolos como potencialmente buenos, o si se quiere, infravalorando la probabilidad de impago de los nuevos clientes. Es por ello que la aplicación de este modelo supondrá aceptar demasiadas veces a potenciales malos clientes (que hubiesen sido rechazados si no existiese sesgo de selección muestral).

Figura 10 Sesgo de selección Muestral en los modelos de *scoring* de admisión



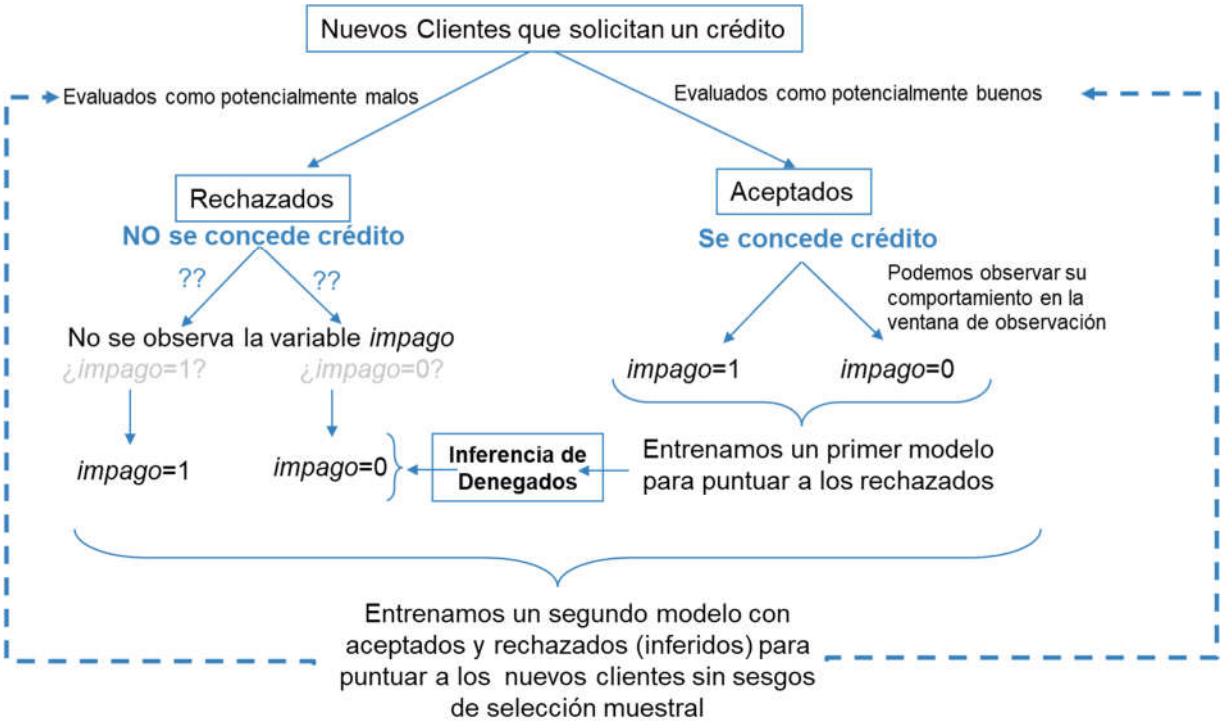
El problema que origina el sesgo de selección radica en que no podemos utilizar a los clientes inicialmente rechazados para construir nuestro modelo por falta de información[1].

[1] Nótese que sólo se evitaría este sesgo de selección muestral en caso de que los rechazados o los aceptados hubiesen sido elegidos inicialmente por la entidad financiera de forma totalmente aleatoria. En ese caso no existiría ninguna relación entre el hecho de ser rechazado o no y la probabilidad de impago. Otro caso en el que tampoco aparece sesgo de selección es en los modelos de *scoring* de comportamiento, en los que se evaluar el riesgo sólo de los ya clientes. En estos modelos se utiliza información de los clientes y se aplica sobre los clientes. La muestra

Si los hubiéramos incluido, entonces nuestra muestra incluiría a aceptados y rechazados y ya sí sería representativa de todos los nuevos clientes solicitantes de un crédito, y por tanto no tendría problema de sesgo de selección alguno.

De hecho, una de las soluciones a este problema del sesgo de selección muestral es precisamente la de inferir de alguna manera qué es lo que hubieran hecho los clientes rechazados en caso de que sí les hubiesen concedido el crédito. Es decir, tenemos que hacer un pronóstico sobre la variable impago de esos clientes rechazados. Una de las técnicas para hacer esa inferencia de denegados consiste en utilizar un primer modelo entrenado sólo con los aceptados para puntuar y pronosticar el impago de los rechazados. De esta forma, una vez que ya se dispone de información completa, tanto de las variables explicativas como de la variable objetivo *impago* tanto para los clientes aceptados como los rechazados, se procedería ya a construir, desde cero (es decir volviendo a realizar todo el proceso de tramificación, agrupación de categorías, selección de variables, estimación y diagnosis del modelo de probabilidad), un nuevo modelo de probabilidad de impago, pero contando ya con aceptados y rechazados Figura 11.

Figura 11 Inferencia de denegados para eliminar el sesgo de selección muestral.



Este nuevo modelo no tendría ya problemas de sesgo de selección muestral y podría ponerse en producción para aplicarse para la evaluación de nuevas solicitudes. Nótese que hemos comentado que este segundo modelo, el definitivo, tendría que volverse a realizar desde cero. Es decir, que para evitar los posibles sesgos arrastrados del modelo

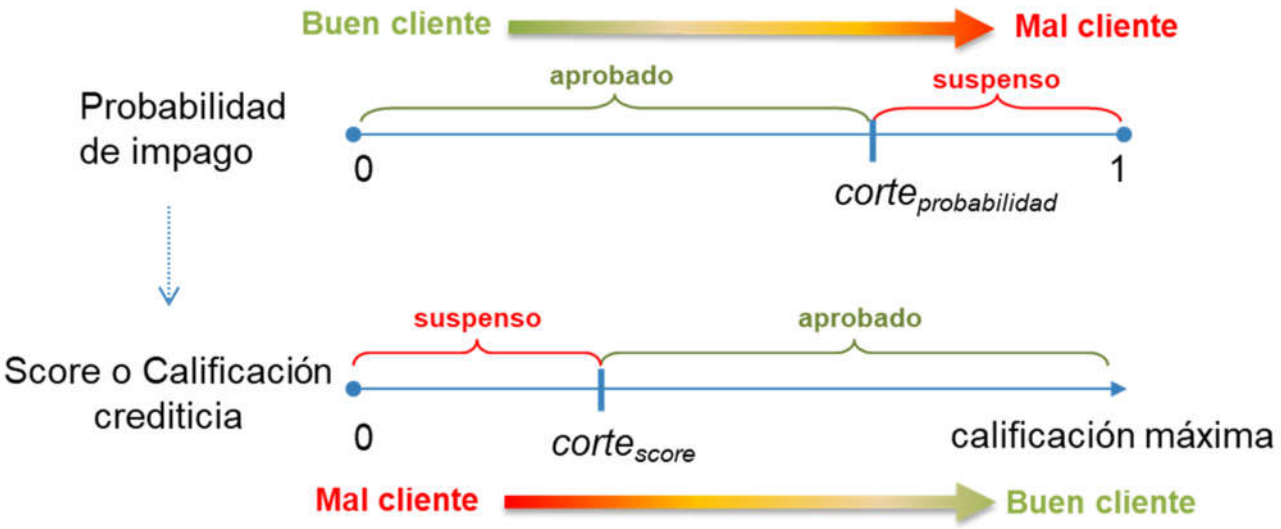
que se utiliza para entrenar y estimar el modelo sí es representativa de la población sobre la que se aplicará el modelo.

inicial habría que realizar, ya con la muestra de clientes aceptados y rechazados (inferidos) todo el proceso de tramificación, agrupación de categorías, selección de variables, transformación WoE, estimación y diagnosis del modelo de probabilidad.

3. De probabilidad de impago a puntuación de la calidad crediticia

Los modelos que hemos comentado anteriormente son modelos de probabilidad, y pronostican para cada nueva solicitud de admisión una probabilidad de impago. Es decir, proporcionan una predicción de la probabilidad de ser mal cliente. En muchos casos, las entidades financieras prefieren trabajar con la denominada puntuación o score de la calidad crediticia, una valoración que es creciente con la calidad crediticia del cliente (menor riesgo de impago). A mayor puntuación mejor valoración se tiene del cliente, y por tano menor probabilidad de que éste incurra en impago (figura 12).

Figura 12. Probabilidad de impago vs Score de calificación crediticia



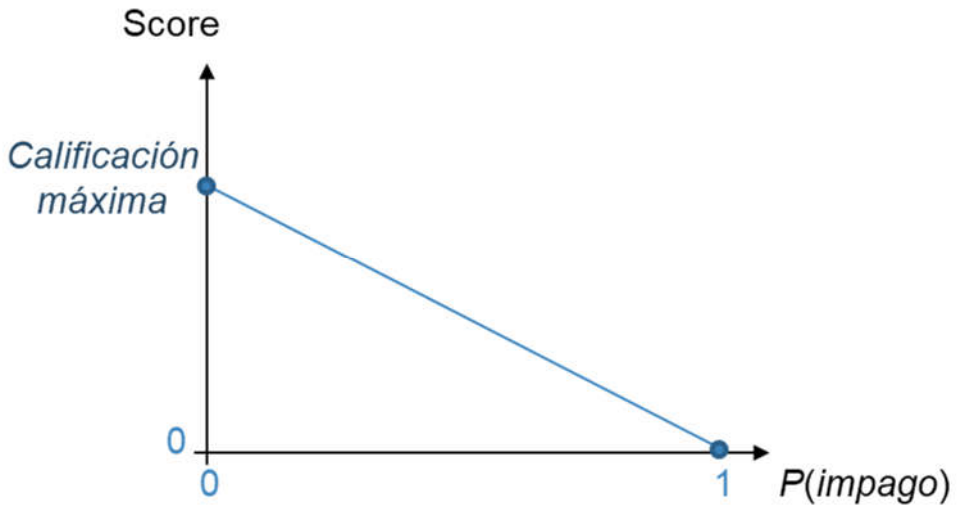
Fíjese que la puntuación o score va en sentido inverso a la probabilidad de impago del cliente. Y su interpretación es como el de la nota que obtenemos en el examen de una asignatura, si la calificación está por debajo de un punto de corte estaríamos suspensos (se nos califica como malos clientes y no se nos concedería el crédito), mientras que superamos ese umbral de corte estaríamos aprobados, y cuanto mayor puntuación mejor nuestra calidad como cliente.

La forma de obtener el score o puntuación es mediante cualquier transformación inversa que convierta probabilidades en puntuaciones, por ejemplo, una transformación lineal como la representada en la figura 13:

$$Score = Calificación\ Máxima\ (1 - probabilidad)$$

Según esta transformación, lo que cada institución debería establecer es cuál es la calificación máxima que quiere otorgar a un cliente con probabilidad de impago P(impago)=0, esto es al mejor de los posibles clientes. También debería establecer el punto de corte para la probabilidad de impago a partir de la cual va a pronosticar que un cliente va a ser un mal pagador y no querrá concederle el crédito. Este punto de corte puede establecerse siguiendo criterios puramente estadísticos (máximo del estadístico *K-S* o máximo del estadístico *F-score*, por ejemplo), o criterios más comerciales (porque cuanto más alto esté este punto de corte más créditos se concederán. A partir de este punto de corte y aplicando la misma relación inversa el punto de corte expresados en términos de puntuación se obtendrá como Calificación Máxima (1- corte en probabilidad). En este caso, las solicitudes de crédito se aprobarán cuando el score o puntuación crediticia del cliente superen este punto de corte de score.

Figura 13. Ejemplo de transformación lineal genérica de Probabilidad de impago en Score de calificación crediticia



3.1 Las tarjetas de puntuación

Un tipo especial de calificación de riesgo crediticio o *scoring* de riesgo, especialmente fácil de interpretar es el que puede realizarse cuando se utiliza como modelo de probabilidad el modelo de regresión logística del que ya hemos hablado anteriormente:

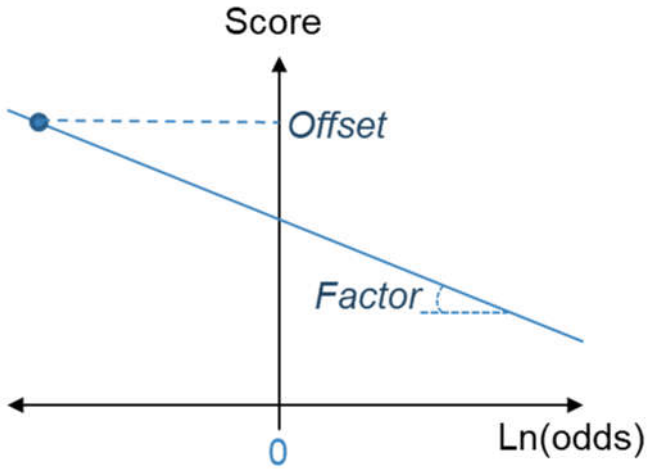
$$P\left(y=1 \mid \boldsymbol{x}^{WoE}\right)=\frac{1}{1-e^{-\left(\beta_0+\beta_1 x_1^{WoE}+\beta_2 x_2^{WoE}+\ldots+\beta_k x_k^{WoE}\right)}}$$

Nótese que, debido a la propia forma funcional de esta función de distribución logística, el logaritmo neperiano del Odd Ratio, definido éste como el cociente entre la probabilidad de impago y la probabilidad de pago tiene una forma lineal en las variables explicativas:

$$\ln\left(\frac{P\,(y=1\mid \boldsymbol{x}^{WoE})}{1-P\,(y=1\mid \boldsymbol{x}^{WoE})}\right)=\beta_0+\beta_1\,x_1^{WoE}+\beta_2 x_2^{WoE}+\ \ldots+\beta_k\,x_k^{WoE}$$

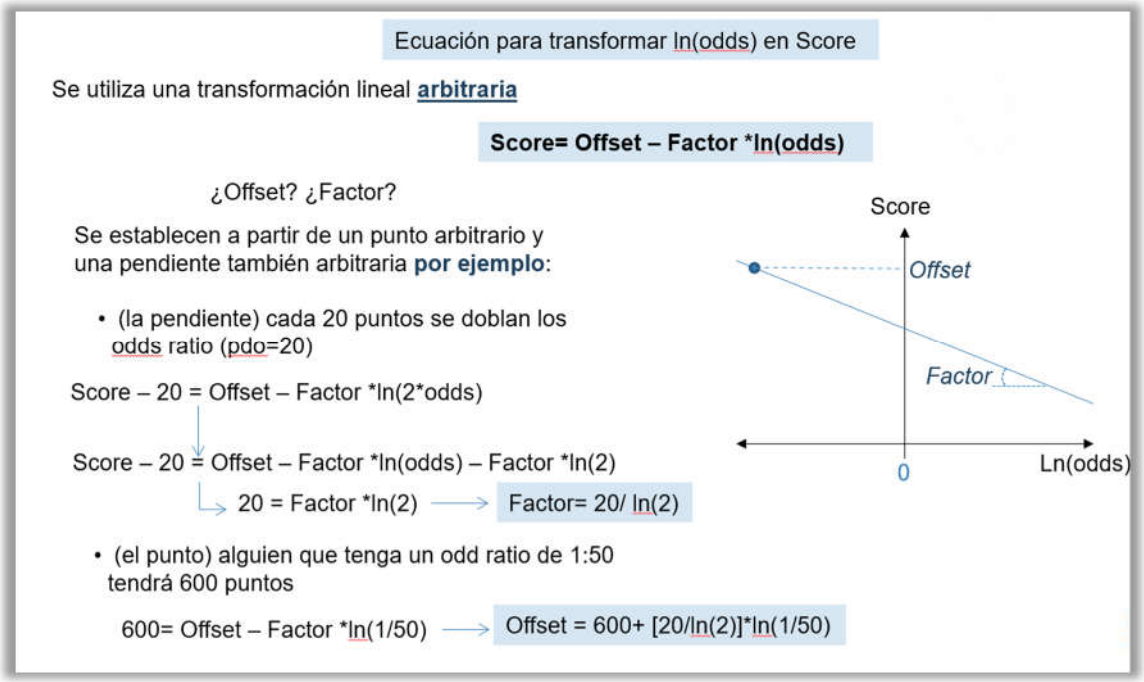
Esta formulación de la regresión logística es conocida como modelo *logit*, y tiene la ventaja de expresar el riesgo de impago (medido por el logaritmo del Odd-Ratio) como una combinación lineal de las variables explicativas, esto es como la suma de diferentes componentes del riesgo, lo que facilita mucho la interpretabilidad de este tipo de modelos.

Figura 14. Transformación lineal entre riesgo de impago y Score de calificación crediticia en el modelo de probabilidad logística



En este caso la transformación desde este riesgo medido por el logaritmo del Odd-Ratio también se realiza a través de una transformación lineal arbitraria, donde además de necesitar de la elección de la calificación que obtendría un cliente muy muy bueno (denominado *Offset*) también es necesario establecer la pendiente de la relación entre Score y riesgo (denominada Factor) tal y como se representa en la figura 14. Ello es debido a que, en este caso, el indicador de riesgo, el logaritmo del Odd-Ratio, no está acotado, como si sucede con la probabilidad, y por tanto para obtener la transformación lineal entre riesgo y Odd-Ratio necesitamos establecer un punto (el Offset) y la pendiente de la recta (Factor):

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



Tanto el punto de corte como la pendiente se eligen de manera arbitraria, aunque normalmente suelen respetarse las escalas que ya estén utilizando previamente cada entidad. Una forma de elegir el Offset (puntuación que obtendría un cliente muy muy bueno) sería, por ejemplo, estableciendo que un cliente con un Odd-Ratio 1:50 (por cada vez que haga impago habría 50 veces que pagaría bien) obtendría 600 puntos. Igualmente, para el Factor o pendiente puede establecerse el número de puntos necesario para doblar el Odd-Ratio, por ejemplo, que cada 20 puntos se doble el Odd-Ratio. A partir de este punto de corte y de esta pendiente sólo queda obtener la expresión para la recta que transforme riesgo (logaritmos de Odd-Ratio) en puntuación. En este ejemplo concreto quedaría:

$$\text{Score} = 600 + \frac{20}{\ln(2)}\left(\ln\left(\frac{1}{50}\right) - \ln(\text{Odds})\right)$$

Como decimos, la ventaja de utilizar la regresión logística como modelo predictivo de la probabilidad de impago es su interpretabilidad. Con la regresión logística es posible expresar el riesgo, medido por el logaritmo del Odd-Ratio, como una combinación lineal de las variables explicativas, y por tanto también es posible expresar su calidad crediticia o scoring como una suma de componentes, cada uno de ellos determinado por el valor que tenga cada individuo en cada una de las variables explicativas. Esta combinación lineal se establece muchas veces en forma de **tarjeta de puntuación o scorecard**, que expresa como deberían puntuarse cada uno de los niveles o categorías de las variables características que integran el modelo.

Figura 15 Scorecard o tarjeta de puntuación de la calidad crediticia

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

De esta forma bastaría con sumar la calificación de cada cliente en cada una de sus características para conocer su calificación o score final, y por tanto aprobar su solicitud de crédito en función de que supere o no el punto de corte establecido por la entidad financiera.

Recordemos que este tipo de tarjeta de puntuación, aunque muy útil por su sencillez de uso, sólo es posible obtenerla cuando se utiliza el modelo de regresión logística como modelo para predecir la probabilidad de impago, ya que es el único que puede transformarse en una combinación lineal (modelo logit). Cada institución debería sopesar en qué medida otro tipo de modelos proporcionan mejores ajustes, aunque sea en detrimento de la obtención de estas tarjetas de puntuación.

Bibliografía básica

- **Puntuación de Riesgo de Crédito**
  - 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
  - Trueck, S, & Rachev, Svetlozar (2009): *Rating Based Modeling of Credit Risk. Theory and Application of Migration Matrices*. Elsevier
- **Modelo de regression logística**
  - Wooldridge, J.M. (2019). *Introductory Econometrics*. A modern Approach. 7th Edition. Cengage Learning Inc. (Cap 17)
- **Modelos de machine Learning**
  - Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.
  - Han, J., Kamber, M., & Pei, J. (2011). *Data mining concepts and techniques* third edition. The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124.
  - Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.
  - Kubat, M. (2017). *An introduction to machine learning*. Springer International Publishing AG.
- **Bibliografía sobre software R para análisis de datos**
  - Dalgaard, P (2008): Introductory Statistcs with R. Springer
  - Wickham, H y Grolemud, G (2017): R for Data Sciece. Visualize, Model, Transform, Tidy and Impor Data" Editorial O'Reilly. Versión actualizada(y gratuita) online https://r4ds.had.co.nz/index.html
  - William, G. (2011). Data Mining with Rattle and R, The art of Excavating Data for Knwledge Discovery. Springer.
- **Aplicación de Riesgos de Crédito con R**
  - Szepannek, G (2022): An Overview on the Landscape of R Packages for Open Source Scorecard Modelling. Risks 2022, 10(3), 67; https://doi.org/10.3390/risks10030067
- **Bibliografía sobre software Python para análisis de datos**
  - Igual, L; y Seguí, S. (2017): Introduction to Data Science. A Python Approach to Concepts, Techniques and Applications. Springer
  - Géron, A. (2022): Hands-On Machine Learning with *scikit-learn*, *Keras* and *TensorFlow*. Concepts, Tools and Techniques to Build Intelligent Systems. Ed. O'Eilly
- **Aplicación de Riesgos de Crédito con Python**
  - Bolder, J. D. (2019). Credit-Risk Modelling: Theoretical Foundations, Diagnostic Tools, Practical Examples, and Numerical Recipes in Python. Springer