



Afi Escuela
de Finanzas

Regresión avanzada

Daniel Vélez Serrano
Febrero 2024

Índice

1. Introducción
2. Técnicas de selección de variables
3. Transformación de variables (I)
4. Modelo de regresión logística
5. Transformación de variables (II): *Weight Of Evidence*
6. Caso de uso: Morosidad
7. Práctica

1 | Introducción

Modelos lineales

- Aun cuando **los modelos lineales suelen moverse dentro del ámbito de la estadística más tradicional**, en los que prima su capacidad explicativa por encima de su capacidad predictiva, a lo largo de los años se ha trabajado en diferentes vías que han llevado a **poder considerar a estos modelos como una técnica más de Machine Learning**, sacrificando parte de dicha capacidad explicativa.
- **Algunas de estas vías** han sido:
 - A** La posibilidad de trabajar con grandes volúmenes de datos: BIGLM, BIGGLM.
 - B** Las técnicas de selección de variables o de estimación de sus parámetros: RIDGE, LASSO, ELASTIC NET.
 - C** Las transformaciones a realizar sobre los datos para mejorar su capacidad **predictiva**: discretización de variables, transformaciones tipo WOE, GAMs.

A Modelos lineales en el ámbito del *Machine Learning*

Adaptación a grandes volúmenes de datos

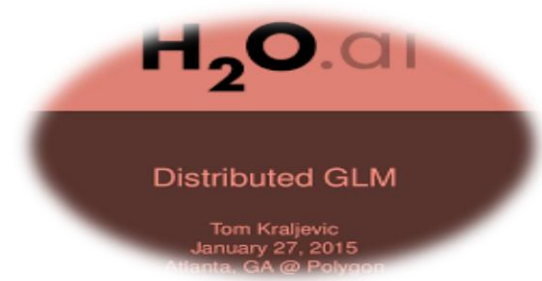
- Cada vez hay **conjuntos de datos más voluminosos** a los que se pretende ajustar un modelo y, aun cuando se ha abaratado el precio de la memoria en los últimos años, **en ocasiones se excede la RAM disponible del ordenador**.
- Si el número de variables (p) se dispara, siempre **se puede recurrir a alguna técnica de reducción de la dimensionalidad**. De hecho, **desde la óptica de los modelos lineales**, hay un tipo de regresión denominada ***Principal Component Regression (PCR)***.
- Sin embargo, **cuando el problema es que hay muchas observaciones (n)**, puede que no sea posible guardar todos los datos en la memoria RAM, existiendo también **desde la óptica de los modelos lineales**, estrategias para **dividir el problema en trozos**, cada uno lo suficientemente pequeño como para caber en la RAM, construyendo gradualmente una solución.



A Modelos lineales en el ámbito del *Machine Learning*

Adaptación a grandes volúmenes de datos

- El paquete **BIGLM** de Thomas Lumley (<http://cran.r-project.org/web/packages/biglm/>), contempla las funciones `biglm()` y `bigglm()`, las cuales permiten respectivamente el ajuste de modelos LM y GLM utilizando una cantidad de memoria del orden $O(p^2)$ en lugar de la habitual $O(n \cdot p^2)$.
- También existe una función implementada bajo el entorno **H₂O** (<https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html>) la cual permite el ajuste de modelos **GLM**, permitiendo aprovechar los diferentes CORES de la máquina (trabajando en paralelo).
- Deber tenerse en cuenta que **los resultados obtenidos por la función GLM de R y H2O no son los mismos** dado que funcionan de manera diferente:
 - **H₂O** usa **H₂O math**, **H₂O objects** y **H₂O distributed computing**.
 - **H₂O GLM** usa términos de regularización (concepto que veremos más adelante), por lo que esencialmente resuelve un problema diferente.



2 | Técnicas de selección de variables

B Técnicas de selección de variables

Métodos RIDGE y LASSO

- La estimación de los parámetros de un modelo de regresión lineal responde a la expresión: $\hat{\beta} = (X'X)^{-1}X'Y (= \frac{cov(X,Y)}{V(X)})$
- Una de las preguntas que surge hace referencia a la **invertibilidad de de $X'X$** .
- Si las variables input presentan un alto nivel de correlación, la matriz **$X'X$ puede no ser invertible** lo que puede derivar en **multicolinealidad**.
- **Ésta no afecta a la predicción dada por el modelo, pero sí a su capacidad explicativa:** los coeficientes no se estiman con errores estándar pequeños (p-valores altos), lo cual no deja de ser un problema parecido al que se tiene cuando se hace un modelo con pocas observaciones.
- Para tratar de solventar este problema, surge la **regresión RIDGE** (Hoerl & Kennard en 1970) como un **método de estimar los coeficientes de la regresión en aquellos casos en los que las variables dependientes están altamente correladas**.

B Técnicas de selección de variables

Métodos RIDGE y LASSO

- La **regresión RIDGE** contempla en el problema de estimación mínimo cuadrático un **parámetro regularizador, llamado parámetro de SHRINKAGE**.

$$\text{minimizar} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^k |\beta_j|^2 = SCE + \lambda \sum_{j=1}^k |\beta_j|^2$$

siendo:

$\lambda = 0 \Rightarrow$ Ajuste Mínimos Cuadrados

- $n = n^\circ$ de observaciones
 - $k = n^\circ$ de variables
 - $\lambda =$ parámetro libre a determinar empíricamente
- $\lambda \uparrow \uparrow \uparrow \Rightarrow$ Modelo constante (todos los parámetros a 0)
(shrink coefficients towards 0)
- El óptimo se alcanza en $\hat{\beta} = (X'X + \lambda I)^{-1} X'Y$. Es la adición del parámetro λ la que **facilita la invertibilidad de $X'X$, y favorece evitar así la multicolinealidad**.
- La **regresión LASSO** (Least Absolute Shrinkage and Selection Operator, Robert Tibshirani, 1996) es una variante de la regresión RIDGE.

$$\text{minimizar} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^k |\beta_j| = SCE + \lambda \sum_{j=1}^k |\beta_j|$$

B Técnicas de selección de variables

Métodos RIDGE y LASSO. Algunas observaciones

- **El valor de los parámetros depende de la escala de medida** de las variables a las que acompaña (variables con rango de valores más altos llevarán asociados parámetros de valor más bajo). Con la idea de que los parámetros (las variables) sean comparables, todas ellas **deben ser previamente estandarizadas**.
- El hacer que los parámetros β_j se vayan hacia 0, juega **en contra del sesgo**. Sin embargo, la **penalización** es **baja** con **respecto a la reducción de varianza**.
- Se puede decir que este problema de optimización consta de dos objetivos: suma de cuadrados debida al error (SCE) y la penalización por parámetros.
- Una forma alternativa de resolverlo es minimizando uno de los objetivos y asociando una restricción al segundo de ellos (método de las e-restricciones).

$$\text{minimizar } \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 \right\} \quad \begin{array}{ll} \text{s.a. } \sum_{j=1}^p |\beta_j|^2 \leq s & \text{RIDGE} \\ \text{s.a. } \sum_{j=1}^p |\beta_j| \leq s & \text{LASSO} \end{array}$$

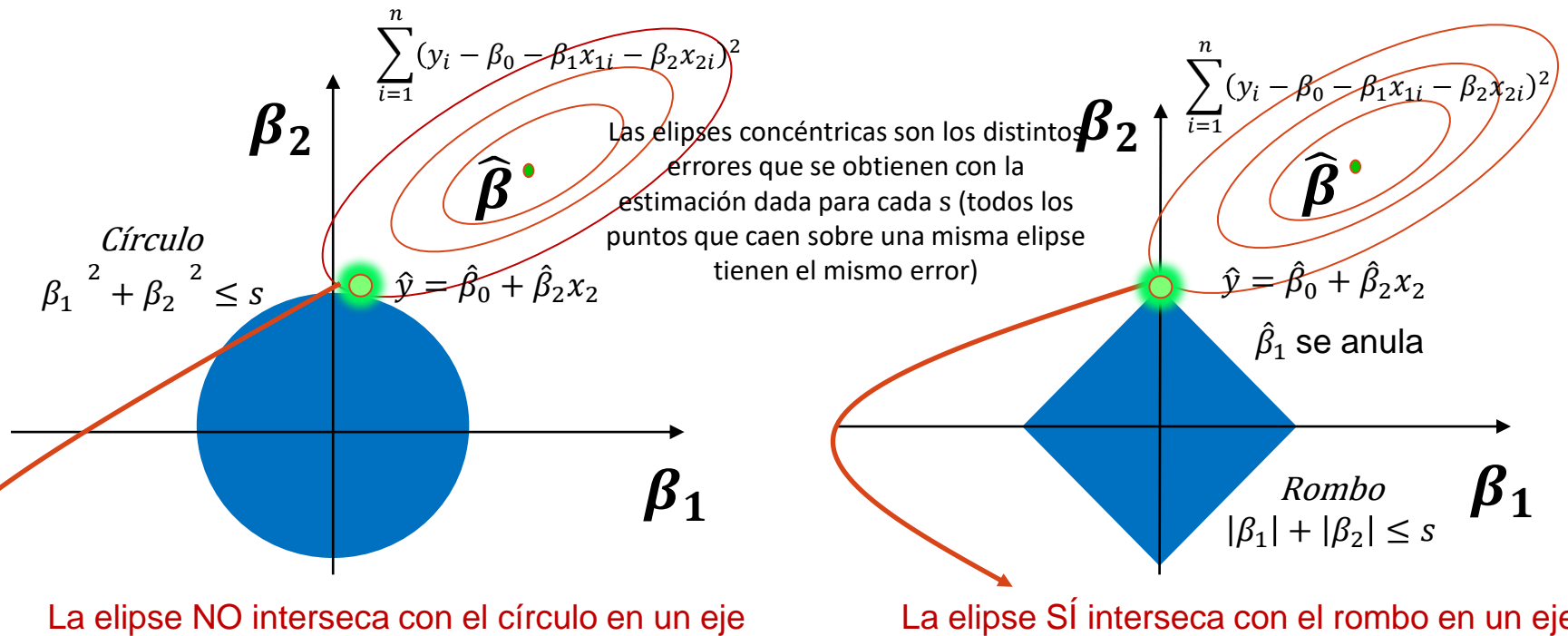
- Si s es lo suficientemente grande, la restricción se relaja y **la estimación de los parámetros es la misma que se obtiene con mínimos cuadrados ordinarios**.

B Técnicas de selección de variables

Métodos RIDGE y LASSO. Algunas observaciones

- **Geométricamente**, supone encontrar el punto de corte entre las distintas soluciones y el espacio de restricciones que se van obteniendo al variar s .

Ej: Supongamos un problema donde únicamente tenemos 2 variables: x_1 y x_2 .



- Intuitivamente, en el gráfico parece que **es más fácilmente que las elipses intersequen en un eje** (lo que lleva a que un parámetro se anule) **en regiones de tipo rombo que en regiones de tipo circular**. Es por ello que **RIDGE no se considera un método de selección de variables y LASSO sí**.

B Técnicas de selección de variables

Métodos RIDGE y LASSO. ELASTIC NET

- Aun cuando **la regresión LASSO es posiblemente la que más tiende a utilizarse**, existe una regla empírica que establece que:

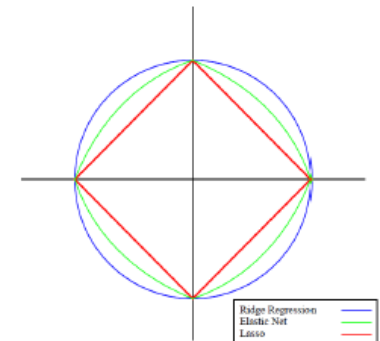
“Si existe un número reducido de regresores que dominan a los demás (con mayor capacidad predictiva), debe usarse la regresión LASSO y, en caso contrario debe usarse la regresión RIDGE”

- En cualquier caso, existen algunas otras variantes intermedias, siendo la más popular la ELASTIC NET, que plantea como objetivo:

$$\text{minimizar } \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^k |\beta_j|^2 + \lambda_2 \sum_{j=1}^k |\beta_j|$$

o equivalentemente:

$$\text{minimizar } \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij})^2 + \lambda((1 - \alpha) \sum_{j=1}^k |\beta_j|^2 + \alpha \sum_{j=1}^k |\beta_j|)$$



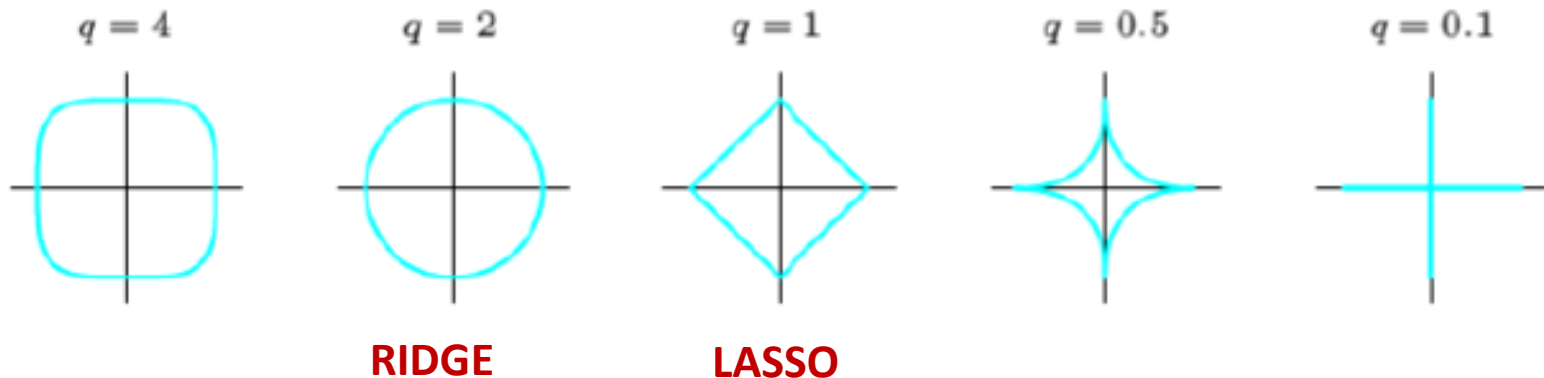
B Técnicas de selección de variables

Métodos RIDGE y LASSO

- También existe **otra versión general** que plantea como objetivo:

$$\text{minimizar } \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^k |\beta_j|^q$$

- En función del **valor de q** , el **espacio** configurado por las **restricciones asociadas a los parámetros** responde a alguna de las siguientes formas:



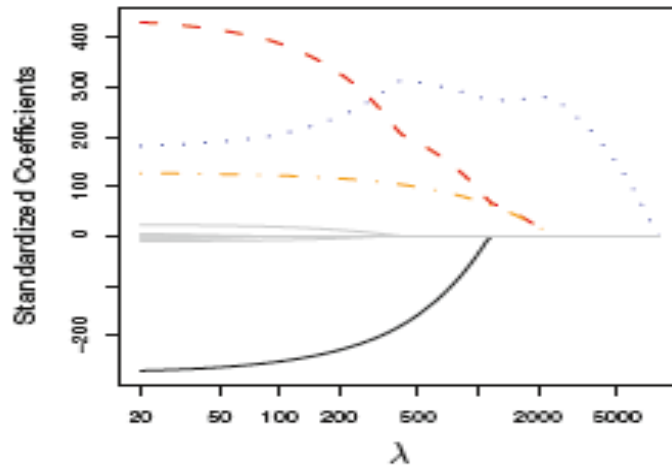
$q > 1 \Rightarrow$ *diferenciable*

$q \leq 1 \Rightarrow$ *métodos más complicados*

B Técnicas de selección de variables

Método LASSO. Resultados

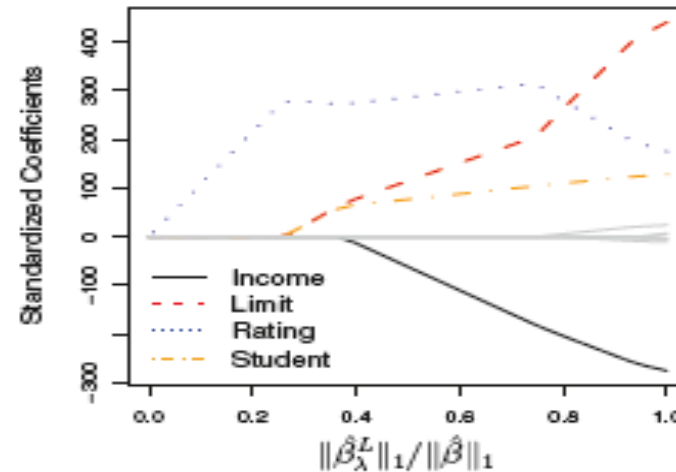
- A diferencia del método MCO que genera solo un conjunto de coeficientes estimados, **la regresión LASSO genera un conjunto de coeficientes estimados β_λ para cada valor de λ .**



Conforme crece λ , los parámetros se van haciendo 0: **“las variables van saliendo del modelo”**.

$\lambda = 0 \Rightarrow$ Ajuste Mínimos Cuadrados

$\lambda \uparrow \uparrow \uparrow \Rightarrow$ Modelo nulo



Una visión alternativa es representar en el eje X el ratio entre el parámetro estimado para un valor de λ y el parámetro estimado por MCO.

Cuando $\lambda = 0 \Rightarrow$ MCO (ratio = 1)

Cuando $\lambda \uparrow \uparrow \uparrow \Rightarrow$ Modelo nulo (ratio = 0)

Gráfico “opuesto”: **“cómo las variables van entrando en el modelo”**.

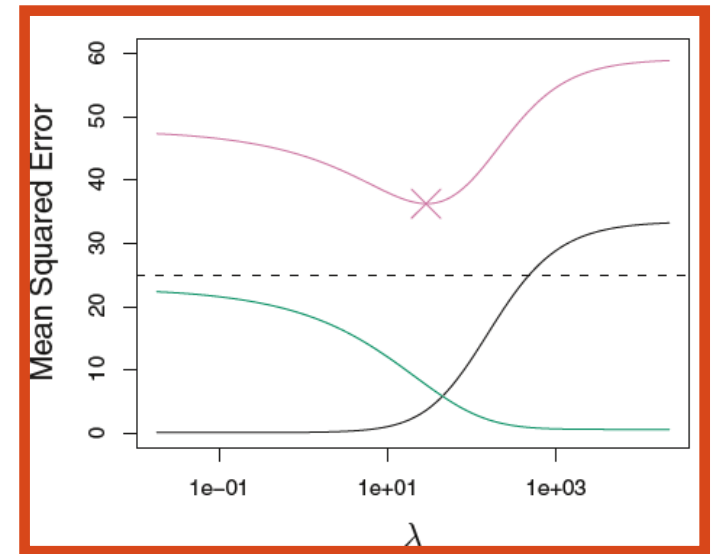
B Técnicas de selección de variables

Método LASSO. Selección del parámetro de *shrinkage*

○ Para determinar el valor de λ :

- Se realiza una batería de pruebas en la que **se varía el valor de λ** .
- Se estiman los parámetros bajo cada uno de esos valores de λ sobre la tabla de train, y **se evalúa el error cometido bajo ese ajuste sobre la tabla de validación**.
- El λ que proporciona el menor error en **validación** será el que dé el valor de los coeficientes estimados.
- **También es habitual** seleccionar el mejor valor de λ por **validación cruzada**.

- La regresión LASSO, **a medida que penaliza y anula coeficientes**, va reduciendo la variabilidad del modelo. Por el contrario, ello implica un **aumento del sesgo** al ajustarse con menor precisión a los datos.



La línea negra representa el sesgo del modelo, la **verde** la varianza, y la **rosa** el error medido sobre un conjunto de test. La **X** marca el punto en el que se obtiene el menor error.

El mínimo se alcanza en un punto donde se consigue una notable reducción de la varianza a cambio de un pequeño aumento del sesgo.

El error obtenido con todas las variables (caso $\lambda = 0$) es mayor.

B Técnicas de selección de variables

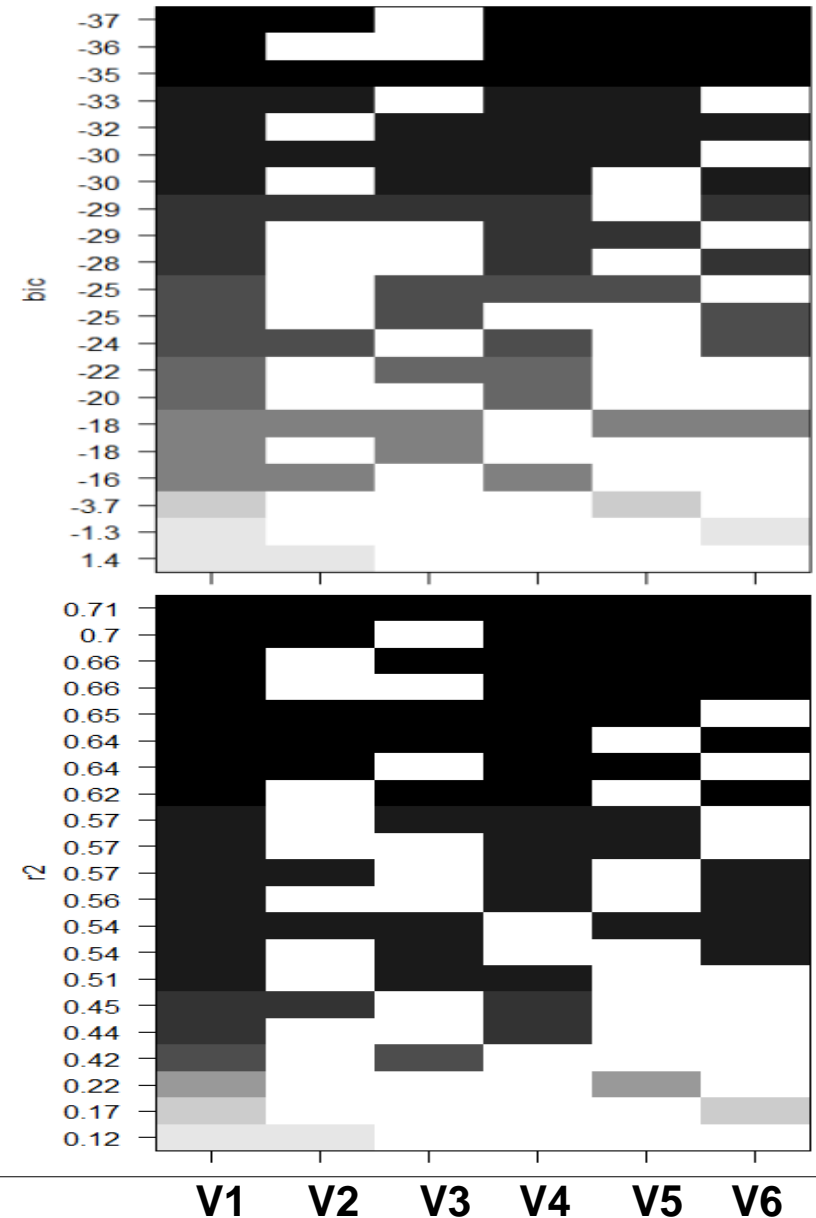
Métodos clásicos

- LASSO, el método más comúnmente utilizado es, como se ha indicado, un **método de selección de variables** frente a otras **clásicas** conocidas como son:
 - **Modelo de selección hacia delante (FORWARD):** empieza seleccionando la variable más correlacionada con la variable respuesta y en cada paso posterior aquella más correlacionada con el error (que mejor explica el error) asociado a cada uno de los modelos que se va obteniendo.
 - **Modelo de selección hacia atrás (BACKWARD):** parte de un modelo que contempla todas las variables explicativas y va eliminando en cada paso, aquella que resulta menos significativa.
 - **Modelo de selección por pasos (STEPWISE):** la estrategia es la misma que el método de selección hacia delante pero además permite la salida de variables del modelo si lo que ésta aporta deja de ser significativo o está explicado (y por tanto redundante) por (con) otras variables contempladas en dicho modelo.
- Una vez establecidas las variables que entran en el modelo, se puede realizar una **validación a posteriori que permite “valorar” cuál de los modelos obtenidos en los diferentes pasos aporta mejor valor respecto de cierta métrica estadística** o de negocio. Con la idea de **evitar el sobreajuste**, esta evaluación suele realizarse sobre una **tabla de validación**.

B Técnicas de selección de variables

Método *Best Subset Selection*

- El **método de selección por pasos** permite **obtener el mejor modelo con una única variable, pero no el mejor modelo con $k (>1)$ variables.**
- El modelo con 2 variables es el mejor modelo con 2 variables, pero condicionando a que la primera de ellas es una variable concreta, no el mejor de los $\binom{p}{i}$ posibles modelos.
- Algunos paquetes estadísticos incorporan un tipo de selección de variables (**best subset selection**) que permite seleccionar el mejor modelo con 1,2,...,k variables.
- Para problemas de minería de datos en los que **el número de ellas (k) puede ser desorbitado, el procedimiento se vuelve inviable: $\sum_{i=1}^k \binom{p}{i}$ modelos**



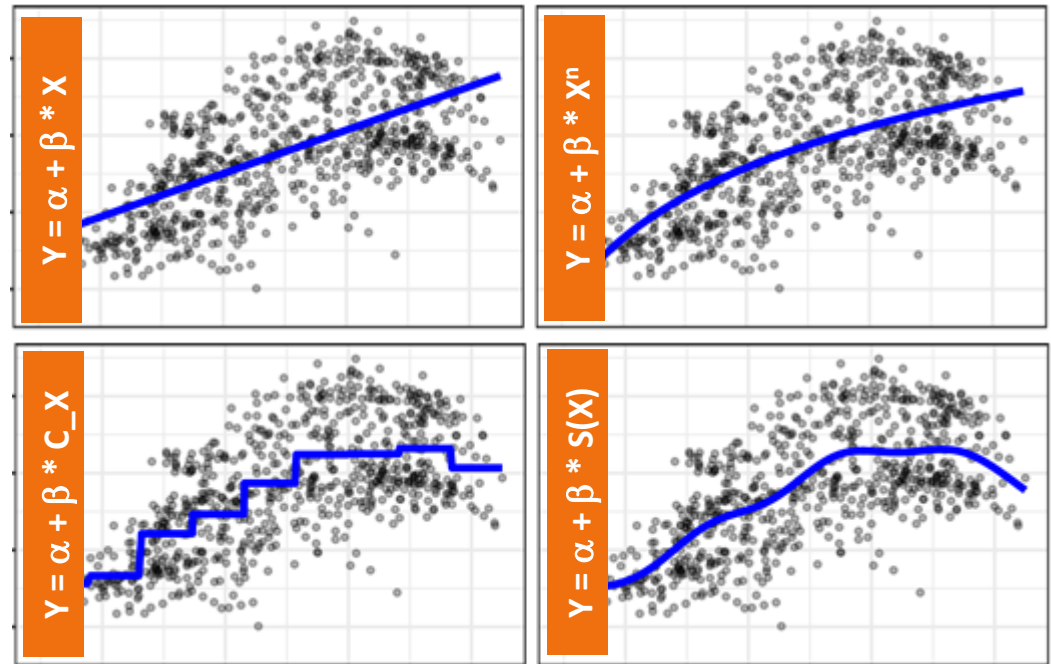
3 | Transformación de variables (I)

C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- Una de las **principales ventajas de las más competitivas técnicas de *Machine Learning*** es su capacidad para **capturar relaciones no lineales entre la variable target y las variables input**. Con vistas a capturar relaciones de este tipo **en un modelo lineal**, **existen diferentes estrategias**:

- **Regresión polinómica**
- **Regresión con variables discretizadas**
- Ajustar **modelos GAMs** (Generalized Additive Models).- en éstos, el término lineal $\beta_j X_j$ es reemplazado por $f_j(X_j)$ donde f_j es una función no lineal “flexible”.



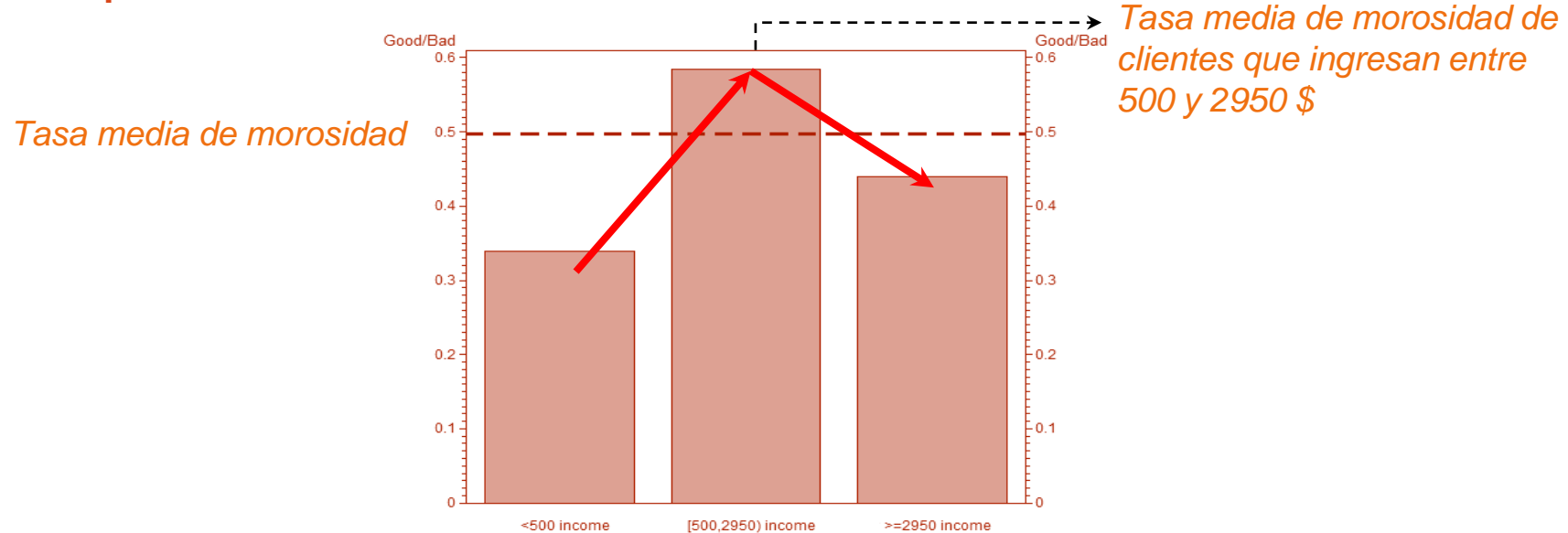
La tipología de funciones que se suele utilizar son **Splines** al constituir funciones que pueden ser combinadas para aproximar cualquier relación.

$$g(E_Y(Y|X)) = \beta_0 + \beta_1 f_1(X_1) + \beta_2 f_2(X_2) + \dots + \beta_p f_p(X_p)$$

C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- La **estrategia de discretización de variables** además de solventar algunos problemas que presentan los modelos lineales como son el efecto de los **outliers** y la imposibilidad de tratar con valores **missings**, permite efectivamente **capturar relaciones de dependencia no lineal**.



- El modelo $Y = \alpha I_{(\dots, 500)} + \beta I_{[500, 2950)} + \gamma I_{[2950, \dots)} + \epsilon$ asigna probabilidades más altas a los clientes que ganan entre 500 y 2950 \$, probabilidades más bajas a los clientes que ganan menos de 500 \$ y probabilidades de valor intermedio a los clientes que ingresan más de 2950 \$.

C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- Además, la **estrategia de discretización** también es un buen recurso cuando la variable de partida presenta **muchas clases**.

Ejemplo:
Variable
Prof

Analysis Variable : GB Good/Bad		
Profession	N Obs	Mean
Chemical Industr	23	0.4782609
Civil Service, M	156	0.3782051
Food,Building,Ca	148	0.5337838
Military Service	28	0.5000000
Others	1428	0.5098039
Pensioner	96	0.3125000
Sea Voyage, Gast	18	0.5000000
Self-employed pe	48	0.7291667
State,Steel Ind,	28	0.5714286



Parameter	Estimate		Standard Error	t Value	Pr > t
Intercept	0.5714285714	B	0.09387135	6.09	<.0001
PROF Chemical Industr	-.0931677019	B	0.13978296	-0.67	0.5052
PROF Civil Service, M	-.1932234432	B	0.10194823	-1.90	0.0582
PROF Food,Building,Ca	-.0376447876	B	0.10236666	-0.37	0.7131
PROF Military Service	-.0714285714	B	0.13275413	-0.54	0.5906
PROF Others	-.0616246499	B	0.09478719	-0.65	0.5157
PROF Pensioner	-.2589285714	B	0.10668621	-2.43	0.0153
PROF Sea Voyage, Gast	-.0714285714	B	0.15006372	-0.48	0.6341
PROF Self-employed pe	0.1577380952	B	0.11811886	1.34	0.1819
PROF State,Steel Ind,	0.0000000000	B	.	.	.

- El hecho de que haya **estimaciones parecidas** entre clases, justificaría la **reagrupación** de clases para favorecer el principio de parsimonia.
- Además, algunas **estimaciones** pueden estar **basadas en soportes bajos** o, simplemente **no resultar significativamente distintas de cero**, lo que justificaría la **necesidad de reagrupar o excluir algunas clases**.

C Modelos lineales en el ámbito del *Machine Learning*

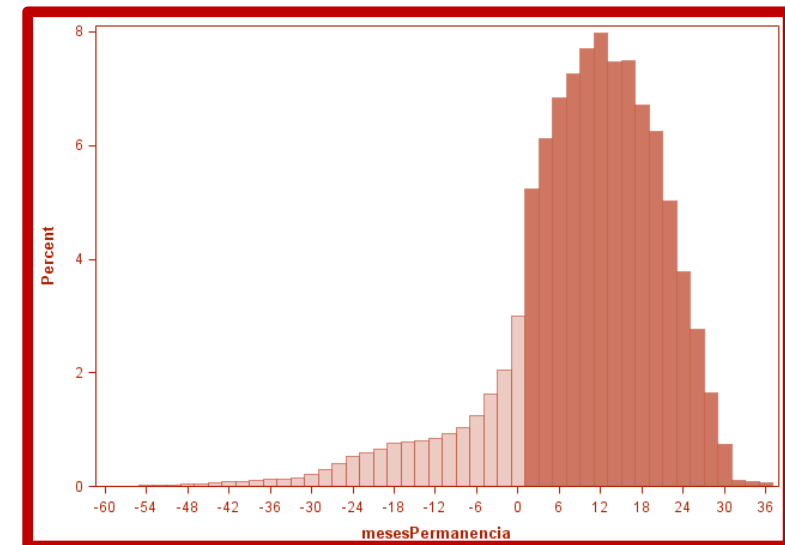
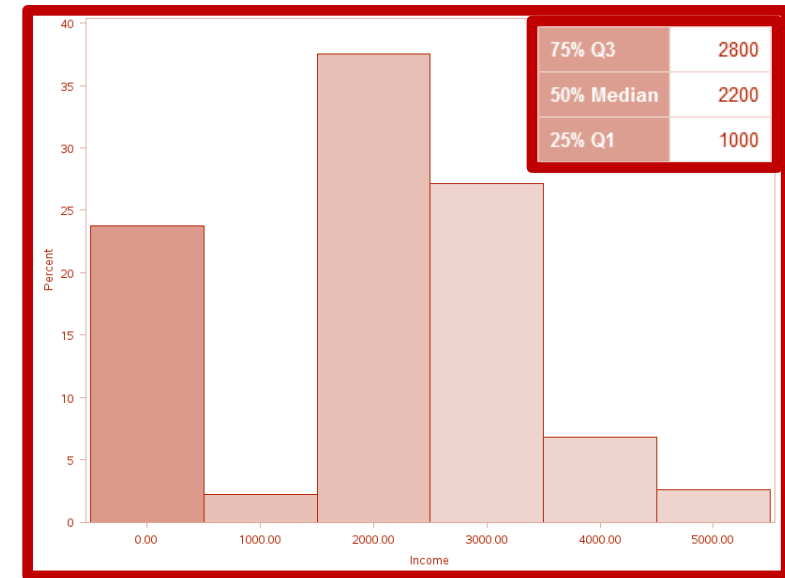
Transformaciones a realizar sobre los datos

- La pregunta es **¿con qué criterio se deben discretizar la variable?:**

1 Una primera posibilidad es discretizar los valores de la variable **de acuerdo a alguna medida de localización de las variables** (cuartiles, deciles, etc.) y asociar al *missing* una categoría especial. **No válido para variables de clase.**

2 Otra posibilidad sería realizar la discretización de acuerdo al **sentido común o de negocio del analista**. Por ejemplo, para un input que refleja el número de meses hasta que expire en contrato de permanencia:

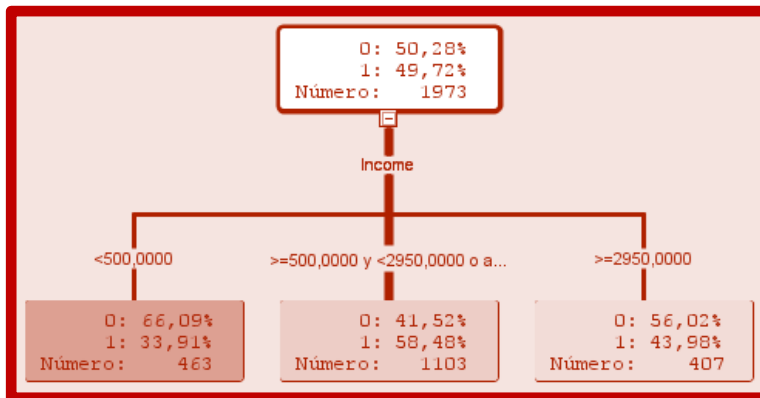
- Clase 1: Valores negativos (contrato de permanencia expirado).
- Clase 2: Valores positivos (incluye el 0) (contrato de permanencia en vigor).



C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- 3 Cuando el analista, no tiene una idea clara de cómo discretizar las variables, lo más conveniente es recurrir a una **técnica analítica**.
- Si la variable fuera de clase una buena estrategia sería agrupar en una misma categoría clases con una tasa de target parecida y diferentes respecto a las otras categorías que se generasen.
 - Si la variable fuera continua, se deberían “tantear” cortes que condujeran a categorías con una tasa de target significativamente distinta.
- Este concepto es precisamente el que subyace bajo el ajuste de modelos **de Árbol de Decisión orientados a discretizar para potenciar el poder predictivo de la variable**.



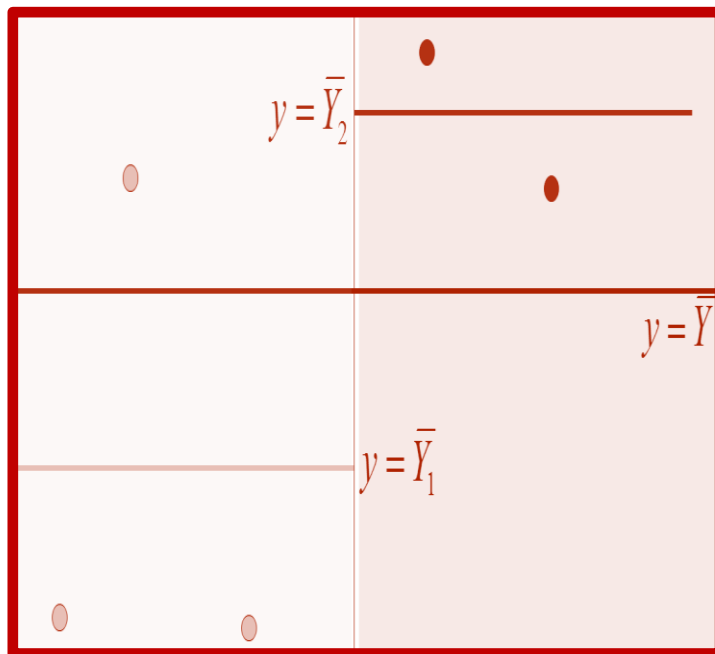
Hay una **no linealidad** en la relación imposible de capturar por la regresión.

Los clientes más propensos a la morosidad son aquéllos que tienen unos ingresos entre 500 y 2950 \$.

C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- En cualquiera de los casos, una vez discretizada la variable, la resultante del proceso **debe ser incluida en el modelo como una variable de clase** (los valores resultantes de la discretización no tienen sentido de orden).
- **La significatividad de dicha variable vendrá dada por el estadístico F asociado al ANOVA** que permite contrastar si dichos valores medios se pueden considerar significativos o no.



Analysis Variable : GB Good/Bad

C_Income	N Obs	Mean
1	463	0.3390929
2	1103	0.5847688
3	407	0.4398034

Tasas de morosidad más alta

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	21.3729487	10.6864743	44.62	<.0001
Error	1970	471.8617193	0.2395237		
Corrected Total	1972	493.2346680			

C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- Obsérvese que, el ajuste de un modelo de regresión lineal para predecir el target en función únicamente de la variable discretizada, proporcionaría unos parámetros que harían referencia al **valor medio** del target en cada clase.

Analysis Variable : GB Good/Bad

C_Income	N	Mean
1	463	0.3390929
2	1103	0.5847688
3	407	0.4398034

⇒

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	0.4398034398 B	0.02425923	18.13	<.0001
C_Income 1	-.1007105672 B	0.03325418	-3.03	0.0025
C_Income 2	0.1449653725 B	0.02838427	5.11	<.0001
C_Income 3	0.0000000000 B	.	.	.

- También se podría generar una **única variable** (no de clase, aunque igualmente binaria) cuyo valor fuese la **media del target dentro de cada clase**.
- Si esta fuese la única variable que participa en el modelo, la estimación del parámetro correspondiente sería obviamente 1 (0 para el intercept), dando lugar a predicciones cuyo valor sería precisamente la media del target en cada clase.

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	0.000000000	0.05376333	0.00	1.0000
w_income	1.000000000	0.10583562	9.45	<.0001

C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- En modelos en los que entran **varias variables** de clase, dichas variables **deben interactuar** entre ellas para que las **predicciones** generadas sean precisamente la **media del valor del target** en cada **cruce de categorías**.
- El hecho de que **no se sepa qué** variables van a acabar realmente **entrando** en el modelo, hace que **carezca de sentido generar cientos de variables** cuyo valor sea la **media del target en posibles cruces** de categorías.
- Salvo que la interacción entre variables resulte clara, lo **habitual es introducir las variables de manera aditiva** (sin multiplicar sus efectos).
- Al convivir varios efectos (variables) en el modelo, la **estimación** de los parámetros no va a ser 1, pero si debería esperarse, al menos en la mayoría de los casos, que fuese positiva (**negativa** en la regresión logística dado que los parámetros aparecen con signos negativos en la fórmula).

$$Y = \frac{1}{1 + e^{-\beta_0 - \beta_1 \text{MediaTargetMesesPermanencia} - \beta_2 \text{MediaTargetIncome} + \varepsilon}}$$

C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- Siguiendo esta línea, **dentro del contexto de modelización de target binario, existe una transformación denominada WOE (Weight Of Evidence, Larsen), muy popular dentro del ámbito del Credit Scoring** en el que, por ley, las entidades bancarias precisan de justificar la no concesión de un crédito a un cliente que lo solicita.

$$\begin{aligned} WOE(X = x_i) &= -\log \left(\frac{\frac{P(Y = 0)}{P(Y = 0|X = x_i)}}{\frac{P(Y = 1)}{P(Y = 1|X = x_i)}} \right) \\ &= -\log \left(\frac{P(Y = 1|X = x_i)}{P(Y = 0|X = x_i)} \right) + \log \left(\frac{P(Y=1)}{P(Y=0)} \right) \propto \log \left(\frac{P(Y = 1|X = x_i)}{P(Y = 0|X = x_i)} \right) \end{aligned}$$

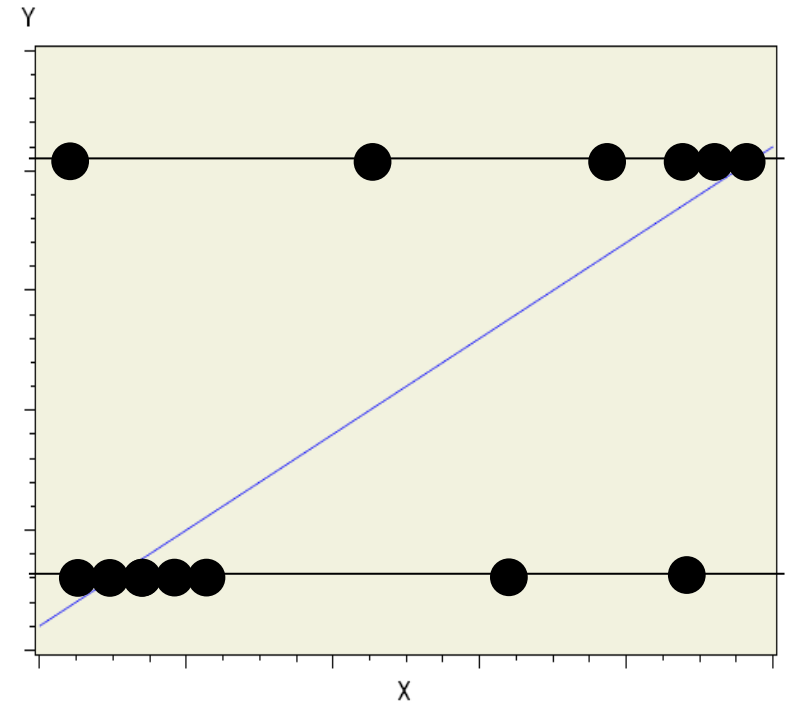
La cantidad $\log \left(\frac{P(Y=1)}{P(Y=0)} \right)$ se suma por igual a todos los nodos que resultan de discretizar y por tanto es prescindible

4 | Modelo de regresión logística

Modelos lineales en el ámbito del *Machine Learning*

Regresión logística

- En el contexto de los **targets binarios**, si se ajustara un modelo de regresión lineal para predecirlo/explicarlo:
 - La variable respuesta es Bernoulli, no normal: contradicción modelo clásico.
 - La varianza de dicha distribución es $p(1-p)$: depende de la media p , no es constante y por tanto presenta heterocedasticidad: contradicción modelo clásico.
 - Los valores estimados se moverían en un rango continuo de la Y , no solo en el intervalo $[0,1]$ (menos aún en $\{0,1\}$): valores no probabilísticos.



Obs: Si el objetivo fue realizar un ranking de clientes (en función de su *scoring* de fuga, morosidad, etc.) el modelo no funcionaría tan mal, aun cuando tampoco sería la mejor solución (no se ajusta bien a pares del tipo $(,0)$ y $(1,)$).

Modelos lineales en el ámbito del Machine Learning

Regresión logística

- Sea $\pi(x) = p(y|X = x) = p(Y = 1|X = x)$ una función de las X 's que define la probabilidad de que la variable Y valga 1 condicionada a una relación de valores para las variables explicativas (X 's).
- Sea entonces $\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right)$, función que puede tomar

cualquier valor real, lo que da sentido a plantear una **regresión lineal** del tipo:

$$\text{logit}(\pi(x)) = \alpha + \beta x + \varepsilon' \quad \text{Se "linealiza" la probabilidad.}$$

$$\frac{\pi(x)}{1 - \pi(x)} = e^{\alpha + \beta x + \varepsilon'} \Leftrightarrow \frac{1 - \pi(x)}{\pi(x)} = e^{-\alpha - \beta x + \varepsilon} \Leftrightarrow$$

$$\text{o equivalentemente: } \Leftrightarrow \frac{1}{\pi(x)} = 1 + e^{-\alpha - \beta x + \varepsilon} \Leftrightarrow \pi(x) = \frac{1}{1 + e^{-\alpha - \beta x + \varepsilon}} \text{ siendo } \varepsilon = -\varepsilon'$$

- Obs: La transformación WOE alude a la propia variable que se predice en la **regresión logística**. Por tanto, se puede ver como el **valor medio del target: target encoding**.

$$\text{WOE}(X = x_i) \propto \log\left(\frac{p(Y = 1|X = x_i)}{p(Y = 0|X = x_i)}\right) = \text{logit}(\pi(x))$$

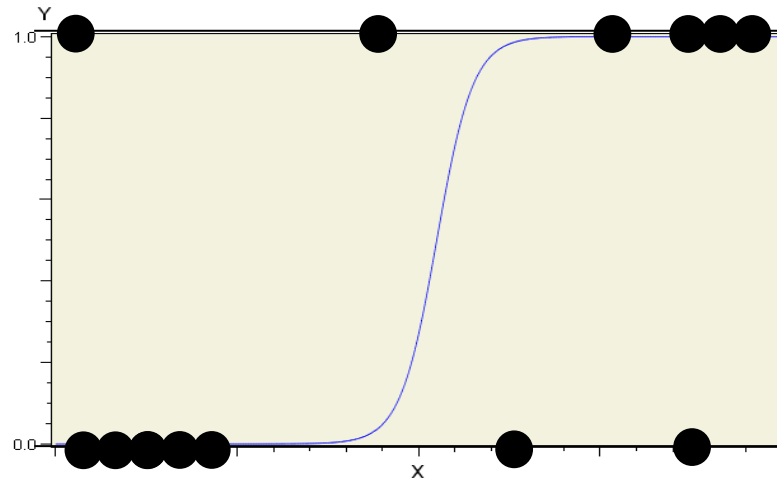
Modelos lineales en el ámbito del *Machine Learning*

Regresión logística

- Este modelo recibe el nombre de **Modelo de Regresión Logística (David Cox, 1958)** que utiliza la función LOGIT como **“función de linkaje”**.

$$Y = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_1 + \dots - \beta_k X_k + \varepsilon}}$$

- Los valores que devuelve se mueven en el rango $[0,1]$ y además, se ajustan bien a pares del tipo $(,0)$ y $(,1)$.



- En general, permite establecer una **relación de dependencia entre una variable dependiente categórica Y (no necesariamente binaria)** con un conjunto de variables independientes de cualquier tipo.

Modelos lineales en el ámbito del *Machine Learning*

Regresión logística

- En función de la naturaleza de la variable dependiente se distinguen diferentes tipos de modelos:
 - **Regresión Logística Binaria:** asociada a un target binario, es la más utilizada y referenciada.
 - **Regresión Logística Ordinal:** asociada a un target ordinal.
 - Existe una ecuación para cada valor del target.
 - Las estimaciones asociadas a los términos constantes de cada ecuación son diferentes.
 - Las estimaciones asociadas a los parámetros de las variables de cada ecuación son las mismas.
 - **Regresión Logística Nominal:** asociada a un target nominal.
 - Existe una ecuación para cada valor del target.
 - Las estimaciones asociadas a los términos constantes y a las variables de cada ecuación (de cada valor del target) son diferentes.

Modelos lineales en el ámbito del *Machine Learning*

Regresión logística

- Además de la función LOGIT, existen otras **funciones de linkaje**:
 - **PROBIT**: inversa de una distribución normal estándar.
 - **cLOG-LOG (o COMPLEMENTARY LOG-LOG FUNCTION)**: que responde a la expresión:

$$\pi(x) = \exp(-\exp(\alpha + \beta x)) \Leftrightarrow \log(-\log(1 - \pi(x))) = \alpha + \beta x$$

- Cuando existen pocos datos los 3 modelos son parecidos, sobre todo el LOGIT y el PROBIT.
- El más utilizado es el LOGIT porque permite realizar una interpretación del valor de sus parámetros a través de los denominados **ODDS RATIOS**.
- Al no ser lineal la relación que plasma la ecuación, no es posible interpretar directamente el valor de los parámetros estimados: “¿cuánto aumenta el valor del target por cada unidad que aumenta el valor de la explicativa?”.

Modelos lineales en el ámbito del *Machine Learning*

Regresión logística

Supóngase X variable predictora, Y variable target y $Z = \beta_0 + \beta_1 X$.

Sea el modelo de regresión logística que estima la probabilidad de Y conocida X .


$$p(y | X = x) = \frac{1}{1 + e^{-z}} \Rightarrow p(y | x) * (1 + e^{-z}) = 1 \Rightarrow p(y | x) * e^{-z} = 1 - p(y | x)$$
$$\Rightarrow e^{-z} = \frac{1 - p(y | x)}{p(y | x)} \Rightarrow \boxed{\frac{p(y | x)}{1 - p(y | x)} = e^z = e^{\beta_0} e^{\beta_1 x}} \quad \text{ODDS}$$

Esta medida refleja, para un valor concreto “ x ” de la variable explicativa X , ¿cuántas veces es más probable obtener valor 1 para el target (ej: padecer una enfermedad, contratar un producto, etc.) que obtener valor 0 (ej: no padecer una enfermedad, no contratar un producto, etc.)?

A través de un ratio de **ODDS** se puede calcular ¿qué influencia genera en el target el incremento de una unidad en el valor de la variable explicativa?

Modelos lineales en el ámbito del *Machine Learning*

Regresión logística


$$OR = \frac{ODDS(X = x + 1)}{ODDS(X = x)} = \frac{e^{\beta_0} e^{\beta_1 (x+1)}}{e^{\beta_0} e^{\beta_1 x}} = e^{\beta_1}$$

ODDS RATIO o
RAZÓN DE ODDS

$$ODDS(X = x + 1) = e^{\beta_1} ODDS(X = x)$$

Así, e^{β_1} es el factor por el cual se multiplica el ODDS de respuesta cuando la variable explicativa se incrementa en una unidad, es decir, ¿cómo se incrementa el ratio de enfrentar la probabilidad del evento frente a la probabilidad del no evento cuando aumenta en una unidad el valor de la variable explicativa?

Es importante destacar que el ODDS Ratio asociado a una variable explicativa es una medida general e independiente del número de variables predictoras: si se mantienen constantes las variables explicativas restantes, sus términos aparecerán tanto en el numerador como en el denominador del cociente y se podrán simplificar.


Modelos lineales en el ámbito del *Machine Learning*

Regresión logística

- **Un valor del ODDS Ratio superior a 1 ($b > 0$), indica que el efecto que produce la variable explicativa X sobre la respuesta Y es de incremento:** aumenta la probabilidad del target.
- **Un valor del ODDS Ratio inferior a 1 ($b < 0$), indica que el efecto que produce la variable explicativa X sobre la respuesta Y es de decremento:** disminuye la probabilidad del target.
- Si la variable explicativa incrementa su valor de “a” a “b” ($b-a$ unidades), la razón de ODDS vale:

$$OR = \frac{ODDS(X = b)}{ODDS(X = a)} = \frac{e^{\beta_0} e^{\beta_1 b}}{e^{\beta_0} e^{\beta_1 a}} = e^{\beta_1 (b-a)}$$

- Ejemplo: Target: Cancer/No Cancer, Explicativa: Fuma/No Fuma.

OR = 2  el hecho de fumar (variable *Fumar* = 1), multiplica por 2 la tasa que refleja cuántas veces es mayor la probabilidad de padecer un cáncer (*Cancer* = 1) que de no padecerlo (*Cancer* = 0).

5 | Transformación de variables (II)

C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- Un valor del ODDS Ratio superior a 1 ($\beta > 0$), indica que el efecto que produce la variable explicativa X sobre la respuesta Y es de incremento: aumenta la probabilidad del target.
- Algunas de las ventajas de las variables tipo WOE son:
- Favorecen decidir si se deja o no una variable en el modelo atendiendo a su “significatividad conjunta” (a un único p-valor en lugar de un p-valor por clase).
 - Las nuevas variables continuas **no presentan outliers** (salvo que los tuviera el target, lo cual no tiene sentido si es binario), **asocian un valor a los missings y recoge no linealidades**.
 - Sus valores han sido obtenidos a través de una **transformación que potencia el poder predictivo de la variable**.
- Además, existe una transformación adicional que permite la conversión del valor **$WOE(X = x_i)$ en una puntuación:**

$$SCORE(X = x_i) = \left(-WOE(X = x_i) * \beta + \frac{\alpha}{n} \right) * factor + \frac{offset}{n}$$

- Dicha transformación es la que reside en la base de los denominados **SCORECARDS**.

C Modelos lineales en el ámbito del Machine Learning

Transformaciones a realizar sobre los datos

- Obsérvese que $WOE(X = x_i) \propto \log \left(\frac{P(Y = 1|X = x_i)}{P(Y = 0|X = x_i)} \right) = \log(ODDS_i)$.
- Así, el **valor WOE** asociado a una clase "i" mide en dicha clase cuántas veces es más probable obtener valor 1 que el valor 0. También **se puede ver como la desviación entre las distribuciones de 1's y 0's en la clase en cuestión con lo cual, cuanto mayor sea su valor, mayor poder discriminante tiene dicha categoría.**
- Es sencillo probar también que $WOE(X = x_i) = \log \left(\frac{P(X=x_i|Y=0)}{P(X=x_i|Y=1)} \right)$, pues:

$$\begin{aligned} WOE_i = WOE(X = x_i) &= -\log \left(\frac{\frac{P(Y = 0)}{P(Y = 0|X = x_i)}}{\frac{P(Y = 1)}{P(Y = 1|X = x_i)}} \right) = -\log \left(\frac{\frac{P(Y = 0)P(X = x_i)}{P(Y = 0 \cap X = x_i)}}{\frac{P(Y = 1)P(X = x_i)}{P(Y = 1 \cap X = x_i)}} \right) = \\ &= -\log \left(\frac{\frac{P(Y = 0)P(X = x_i)}{P(X = x_i|Y = 0)P(Y = 0)}}{\frac{P(Y = 1)P(X = x_i)}{P(X = x_i|Y = 1)P(Y = 1)}} \right) = -\log \left(\frac{P(X = x_i|Y = 1)}{P(X = x_i|Y = 0)} \right) = \log \left(\frac{P(X = x_i|Y = 0)}{P(X = x_i|Y = 1)} \right) \end{aligned}$$

C Modelos lineales en el ámbito del *Machine Learning*

Transformaciones a realizar sobre los datos

- A partir de estos valores WOE se definen los **coeficientes IV (Information Value)** como una **métrica de referencia que mide el poder predictivo de la discretización realizada sobre X para predecir la variable Y.**

$$IV = \sum_i (P(X = x_i|Y = 0) - P(X = x_i|Y = 1))WOE_i = \sum_i (P(X = x_i|Y = 0) - P(X = x_i|Y = 1)) \log \left(\frac{P(X=x_i|Y=0)}{P(X=x_i|Y=1)} \right) = \sum_i (P(X = x_i|Y = 0) - P(X = x_i|Y = 1)) (\log(P(X = x_i|Y = 0)) - \log(P(X = x_i|Y = 1)))$$

- En dicha métrica **se tiene en cuenta la importancia de cada una de las desviaciones** entre las distribuciones de 1's y 0's dadas por las variables WOES. Dicha importancia es cuantificada a través de $P(X = x_i|Y = 0) - P(X = x_i|Y = 1)$.
- De esta forma, un ODDS de 0,02/0,01 es menos importante (0,02-0,01 = 0,01) que uno de 0,2/0,1 (0,2-0,1 = 0,1) aun cuando ambos ODDS son los mismos.
- Obsérvese por ejemplo que en el caso en el que la tasa de 1's y 0's sea la misma (0,5), tanto las importancias (0,5 - 0,5) como las desviaciones ($\log(0,5/0,5) = \log(1)$) tomarían el valor 0, por lo que $IV = 0$.

C Modelos lineales en el ámbito del Machine Learning

Transformaciones a realizar sobre los datos

Tabla de x por y			
x	y		Total
	0	1	
1	2	1	3
	25.00	12.50	37.50
	66.67	33.33	
	40.00	33.33	
2	1	1	2
	12.50	12.50	25.00
	50.00	50.00	
	20.00	33.33	
3	2	1	3
	25.00	12.50	37.50
	66.67	33.33	
	40.00	33.33	
Total	5	3	8
	62.50	37.50	100.00

a(i)

P(X = Y =)	0	1	P(X=x(i) Y=0) - P(X=x(i) Y=1)
1	0,40	0,33	0,07
2	0,20	0,33	-0,13
3	0,40	0,33	0,07

b(i)

ln[P(X = x(i) Y = 0)]	0	1	ln[P(X=x(i) Y=0)] - ln[P(X=x(i) Y=1)]
1	-0,92	-1,10	0,18
2	-1,61	-1,10	-0,51
3	-0,92	-1,10	0,18

$$\sum a(i) * b(i)$$



0,01215477
0,06811008
0,01215477

$$IV = 0,924196$$

Siddiqi (2005) propuso unas reglas para evaluar la bondad de un predictor discretizado X en función del valor de su IV



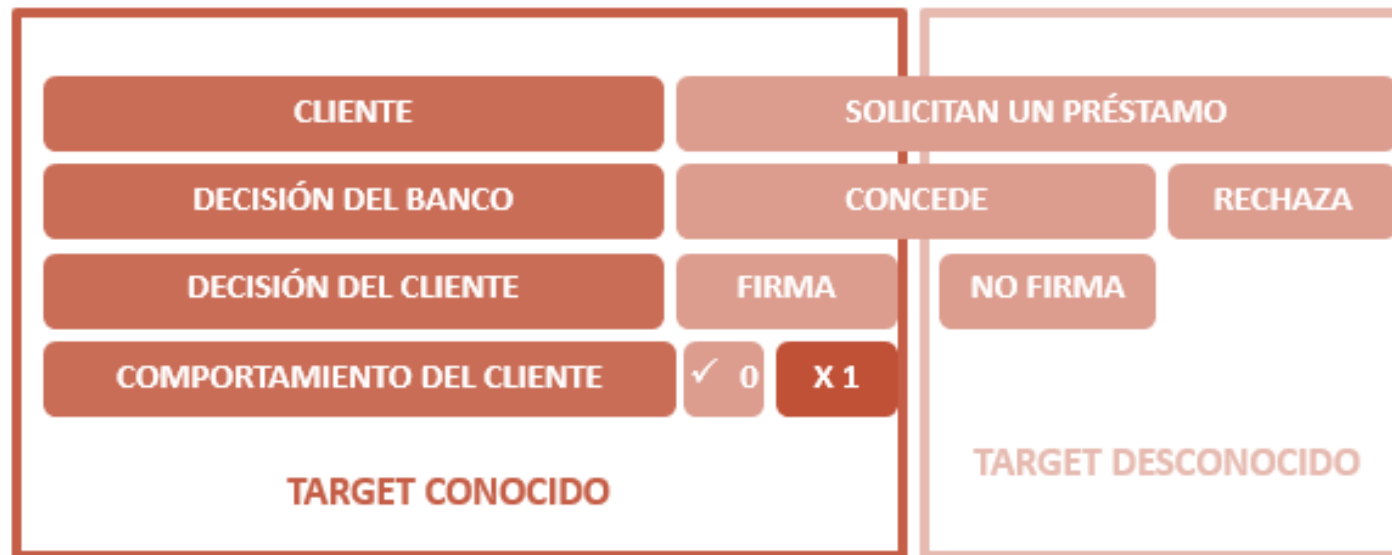
- < 0,02 No útil para predicción
- <0,1 Poder predictivo bajo
- <0,3 Poder predictivo medio
- <0,5 Poder predictivo alto
- >0,5 Poder predictivo sospechosamente alto

6

Caso de uso: Morosidad

Caso de uso. Morosidad

- Las entidades bancarias disponen de una relación de reglas en función de las cuales deciden conceder o no préstamos personales a aquellos posibles clientes que los solicitan. En consecuencia, la información que queda almacenada en sus sistemas no está siempre completa.
- Por esta razón, la naturaleza de los **modelos** a utilizar para ayudar al banco a identificar patrones morosos tienen un carácter **semi-supervisado**.



Caso de uso. Morosidad. Ejemplo: German Data

- Se utilizarán los datos descargados de:

<https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29>.

- La **variable target** se denomina *creditability* y toma los valores 1 (GOOD) y 0 (BAD) en una proporción de 67% - 33%.
- Dado que es **habitual** que se **asigne el valor 1 a la clase más difícil de predecir** (en esta caso, el hecho de “ser moroso”), se realizará una **transformación** al inicio que **intercambie los valores del target**.

```
#####  
# El valor 1 va asociado a la clase GOOD (no moroso) y el 0 a la clase BAD (moroso) #  
# Se asocia el valor 1 a la clase más difícil de predecir: "ser moroso" #  
#####  
  
german_data$creditability=1-german_data$creditability
```

- Se pide: **Ajustar un modelo de regresión logística con variables tipo WOE.**

Caso de uso. German Data. Inputs

Attribute 1: (qualitative)

Status of existing checking account

A11 : ... < 0 DM

A12 : 0 <= ... < 200 DM

A13 : ... >= 200 DM / salary assignments for at least 1 year

A14 : no checking account

Attribute 2: (numerical)

Duration in month

Attribute 3: (qualitative)

Credit history

A30 : no credits taken/ all credits paid back duly

A31 : all credits at this bank paid back duly

A32 : existing credits paid back duly till now

A33 : delay in paying off in the past

A34 : critical account/ other credits existing (not at this bank)

Attribute 4: (qualitative)

Purpose

A40 : car (new)

A41 : car (used)

A42 : furniture/equipment

A43 : radio/television

A44 : domestic appliances

A45 : repairs

A46 : education

A47 : (vacation - does not exist?)

A48 : retraining

A49 : business

A410 : others

Attribute 5: (numerical)

Credit amount

Attribute 6: (qualitative)

Savings account/bonds

A61 : ... < 100 DM

A62 : 100 <= ... < 500 DM

A63 : 500 <= ... < 1000 DM

A64 : .. >= 1000 DM

A65 : unknown/ no savings account

Attribute 7: (qualitative)

Present employment since

A71 : unemployed

A72 : ... < 1 year

A73 : 1 <= ... < 4 years

A74 : 4 <= ... < 7 years

A75 : .. >= 7 years

Attribute 8: (numerical)

Installment rate in percentage of disposable income

Attribute 9: (qualitative)

Personal status and sex

A91 : male : divorced/separated

A92 : female : divorced/separated/married

A93 : male : single

A94 : male : married/widowed

A95 : female : single

Attribute 10: (qualitative)

Other debtors / guarantors

A101 : none

A102 : co-applicant

A103 : guarantor

Attribute 11: (numerical)

Present residence since

Attribute 12: (qualitative)

Property

A121 : real estate

A122 : if not A121 : building society savings agreement/

A123 : if not A121/A122 : car or other, not in attribute 6

A124 : unknown / no property

Attribute 13: (numerical)

Age in years

Attribute 14: (qualitative)

Other installment plans

A141 : bank

A142 : stores

A143 : none

Attribute 15: (qualitative)

Housing

A151 : rent

A152 : own

A153 : for free

Attribute 16: (numerical)

Number of existing credits at this bank

Attribute 17: (qualitative)

Job

A171 : unemployed/ unskilled - non-resident

A172 : unskilled - resident

A173 : skilled employee / official

A174 : management/ self-employed/
highly qualified employee/ officer

Attribute 18: (numerical)

Number of people being liable to provide maintenance for

Attribute 19: (qualitative)

Telephone

A191 : none

A192 : yes, registered under the customers name

Attribute 20: (qualitative)

foreign worker

A201 : yes

A202 : no

7 | Práctica

Práctica. Predicción de morosidad + inferencia de rechazados

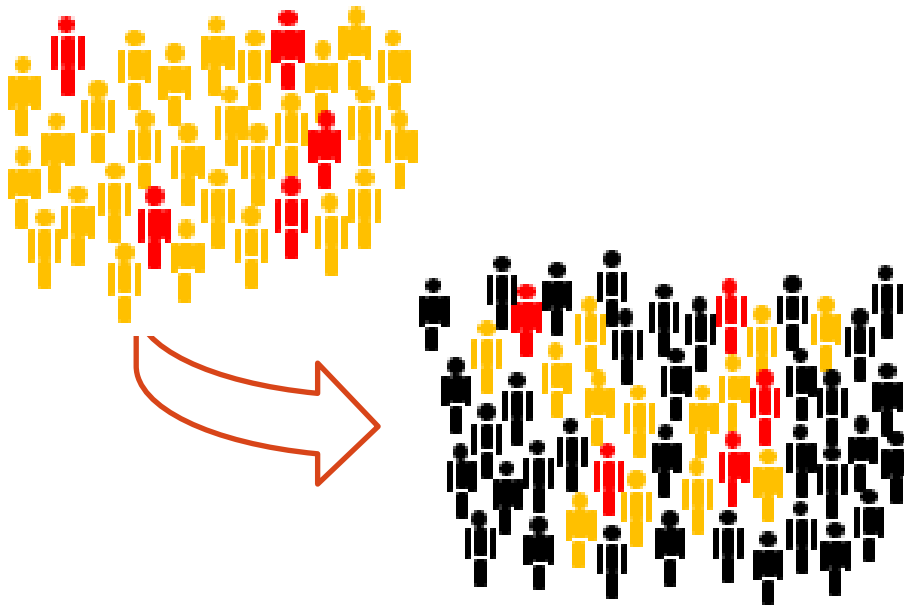
- El **objetivo** es **ajustar un modelo que prediga la probabilidad de ser moroso** de un cliente que solicita un crédito. A tal fin se proporcionan las siguientes tablas:
 - **ACEPTADOS_TRAIN/TEST.**- contiene las solicitudes de crédito que han sido aceptadas y para las cuales se conoce si el cliente terminó devolviéndolo (0) o no (1, fue moroso). Para dichas solicitudes se dispone del valor de las variables explicativas en el momento de hacer la solicitud. En ambas tablas se ha hecho un balanceo de clases 50-50 (la tasa real es del 5%).
 - **RECHAZADOS.**- contiene las solicitudes de crédito que fueron rechazadas y para las cuales no se conoce por tanto si el cliente fue o no moroso. Para dichas solicitudes solo se dispone del valor de las variables explicativas en el momento de hacer la solicitud.
- Se pide:
 - Ajustar un modelo (modelo 1) con la tabla **ACEPTADOS_TRAIN** y ver su capacidad de acierto en la tabla **ACEPTADOS_TEST**.
 - De manera adicional (modelo 2), se deberá utilizar la tabla **RECHAZADOS** para tratar de mejorar la capacidad de acierto del modelo ajustado sobre **ACEPTADOS_TEST**, utilizando a tal fin alguna técnica de inferencia de rechazo.
- Presentar un .html/.pdf en el que se Justifiquen las decisiones tomadas durante el proceso de modelización(tratamiento de outliers/missings, método de selección de variables, etc.) e interpretar los resultados del modelo (realizar la evaluación en término del valor del *lift* en el centil 5 con ambos modelos).

Práctica. Predicción de morosidad + inferencia de rechazados

Variable	Descripción
AGE	Edad
BUREAU	Clase de riesgo de la Oficina de Crédito
CAR	Tipo de vehículo
CARDS	Tipo de tarjeta de crédito (Visa, MasterCard,...)
CASH	Dinero en efectivo solicitado
CHILDREN	Número de niños
DIV	Región grande (1) o no (0)
EC_CARD	Titular de tarjetas EC
FINLOAN	Número de préstamos terminados
GB	Si ha pagado (<u>Good=0</u>) o no (<u>Bad=1</u>)
INC	Salario
INC1	<u>Salario+EC Card</u>
INCOME	Ingresos
LOANS	Número de préstamos en curso
LOCATION	Oficina de crédito con localización física (1) o no (0)
NAT	Nacionalidad
NMBLOAN	Número de préstamos Mybank
PERS_H	Número de habitantes en el hogar
PRODUCT	Tipo de negocio
PROF	Profesión
REGN	Región
RESID	Tipo de residencia
STATUS	Estado
TEL	Teléfono
TITLE	Título
TMADD	Tiempo en casa
TMJOB1	Tiempo en el trabajo

Técnicas de inferencia de rechazados

- La **muestra** con la que el modelo se entrena **presenta sesgo**: está conformada por **clientes que no son sospechosos de ser morosos** por no haber hecho saltar ninguna de las reglas de alerta de la compañía.



Técnicas de inferencia de rechazados

- Dichas alertas saltan de acuerdo al valor conocido o declarado de las variables que definen dicho perfil (ingresos, formar parte de la lista de ASNEF (*), etc.).
- Se dispone así de **clientes para los que se conoce el valor de las variables explicativas pero no el de su target**: el crédito no es concedido y no existe la posibilidad de saber si su comportamiento de pago derivaría o no en morosidad.
- La **pregunta** que se plantea es **si es posible aprovechar el valor registrado de dichos inputs** para mejorar la calidad predictiva del modelo ajustado.
- La solución pasa por **inferir** el target para esas **solicitudes rechazadas**.

(*) ASNEF.- Asociación Nacional de Establecimientos Financieros de Crédito.- fundamentalmente conocida por su registro de morosos (fichero ASNEF), que contiene información sobre todas las personas que tienen una deuda impagada con alguno de sus socios

Técnicas de inferencia de rechazados

- Existen diferentes métodos para llevar a cabo la inferencia de rechazados:

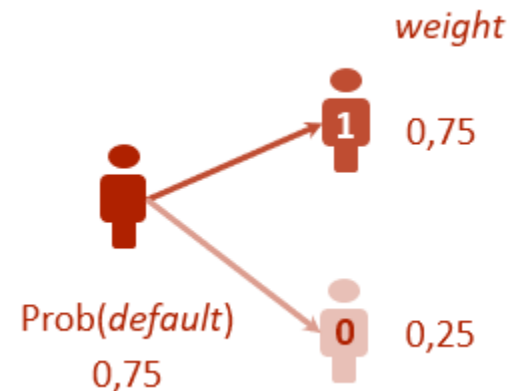
- **Hard cutoff.-**

- Se puntúa con el modelo ajustado a toda la población rechazada.
- Se **establece un valor probabilístico** por encima del cual el target será considerado 1 y por debajo del cual será considerado 0.
- Se ajusta un modelo que se entrene tanto con la muestra de clientes aceptados como con los rechazados con target inferido.

- **Fuzzy method.-**

- Se puntúa con el modelo ajustado a toda la población rechazada.
- **Por cada observación de dicha población se generan dos observaciones: una a la que se asigna target 1 (morosidad) y otra a la que se asigna target 0 (no morosidad).**
- A cada una de dichas observaciones se asigna un **peso proporcional a la probabilidad** de la clase en cuestión.
- Se ajusta un modelo que se entrene tanto con la muestra de solicitudes aceptadas como con la muestra rechazada ponderada.

Id	Pr	T
3	0.9	1
2	0.7	1
4	0.3	1
5	0.1	0
1	0.05	0



Técnicas de inferencia de rechazados

- Existen diferentes métodos para llevar a cabo la inferencia de rechazados:

- **Parceling.-**

- Se puntúa con el modelo ajustado a toda la población rechazada.
- Se **distribuye la muestra de clientes rechazados en intervalos** definidos por rango de score elegidos: de igual longitud, scores distintos (cruces de variables de grupo, WOE), etc.
- Los **rechazos** dentro del intervalo son **clasificados al azar como evento o no evento**, teniendo en cuenta la **probabilidad del modelo en los intervalos (medias si las probabilidades varían dentro de dichos intervalos)**.

Score	#Bad	#Good	%Bad	%Good	Reject	Rej-Bad	Rej-Good
0-169	290	971	23%	77%	1.646	379	1.267
170-179	530	2.414	18%	82%	1.732	312	1.420
180-189	365	2.242	14%	86%	3.719	521	3.198
...
230-239	139	6.811	2%	98%	3.871	77	3.794
240-249	88	10.912	0.8%	99.2%	4.773	38	4.735
250 +	94	18.705	0.5%	99.5%	8.982	45	8.937

Técnicas de inferencia de rechazados

- Existen diferentes métodos para llevar a cabo la inferencia de rechazados:

- **Augmentation.-**

- **El modelo sirve realmente para seleccionar, dentro de las solicitudes que serán aceptadas, cuáles realmente incurrirían en morosidad.** Dado que el modelo funciona bien bajo estas condiciones se podría solicitar a la compañía cuáles son sus criterios de aceptación de solicitudes y aplicar el modelo sobre aquéllas que van a ser aceptadas.
- **Inconveniente:** los criterios pueden no ser claros, no ser uniformes o incluso la compañía puede ser reacia a proporcionarlos por querer un modelo que detecte nuevos patrones.
- **Solución:**
 - **Ajustar un modelo para predecir las reglas de morosidad de la compañía.** Se entrenará con solicitudes aceptadas ($\text{target} = 1$) y rechazadas ($\text{target} = 0$). El modelo proporcionará la probabilidad de que una solicitud sea aceptada.
 - **Puntuar con dicho modelo las solicitudes aceptadas** (para las que se tiene el target de morosidad).
 - Asignar a **cada solicitud aceptada un peso inversamente proporcional a la probabilidad de aceptación** asignada por el modelo anterior.
 - Ajustar el **modelo de morosidad** sobre solicitudes aceptadas teniendo en cuenta dichos pesos: **observaciones con baja probabilidad de ser aceptadas tendrán más peso en el modelo.**



Afi Escuela
de Finanzas

danielvelezserrano@gmail.com