

**Fastbook 08**

**Estadística Aplicada  
al Marketing**

**Regresión lineal y regresión logística**



## 08. Regresión lineal y regresión logística

En el fastbook anterior descubrimos:

- **Distribuciones y función de densidad y distribución.** El concepto de distribución (modelización del comportamiento de una variable aleatoria) basado en el comportamiento de la función de distribución (probabilidad acumulada) y la función de densidad o probabilidad, según sea continua o discreta. Conocimos distintos tipos de distribuciones continuas y discretas, centrándonos en la distribución normal.
- **Intervalos de confianza.** Descubrimos la gran y extensa utilidad de los intervalos de confianza basados en el nivel de confianza  $\alpha$  y en el comportamiento de la variable (distribución a la que se asemeja: normal, t-student, etc.).
- **Contraste de hipótesis - ANOVA.** Por último, conocimos el significado y utilidad de los contrastes de hipótesis: técnicas estadísticas para estudiar la relación entre dos o más variables una de ellas categórica. Centramos la atención en la técnica más utilizada: el ANOVA, que estudia el comportamiento de dos o más grupos de poblaciones basados en la media.

En este fastbook conoceremos el modelo matemático más simple a la par que extendido: la regresión (lineal y logística). Conoceremos su comportamiento, cómo modelizarlo, su utilidad y cómo leer este tipo de modelos (coeficientes, p-valores, ajuste).

*Autora: Patricia Martín González*

≡ Antes de empezar

≡ Regresión lineal

≡ Regresión lineal simple

≡ Regresión lineal múltiple

≡ Regresión logística

≡ Resumen

# Antes de empezar

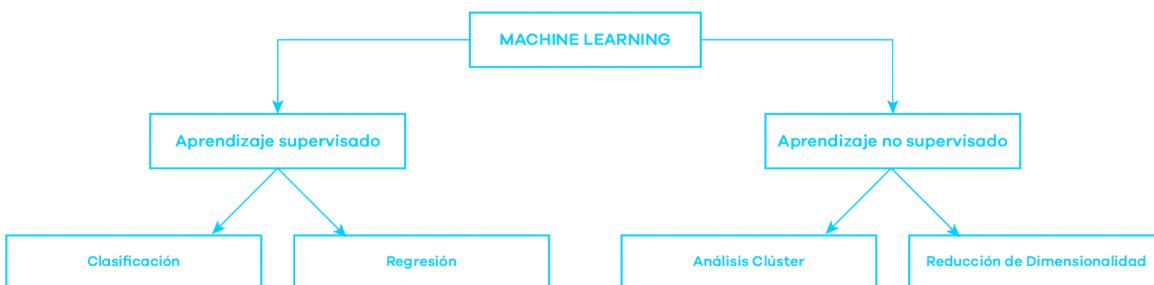
X Edix Educación

Una gran variedad de modelos matemáticos busca explicar una variable objetivo en función de otras variables:

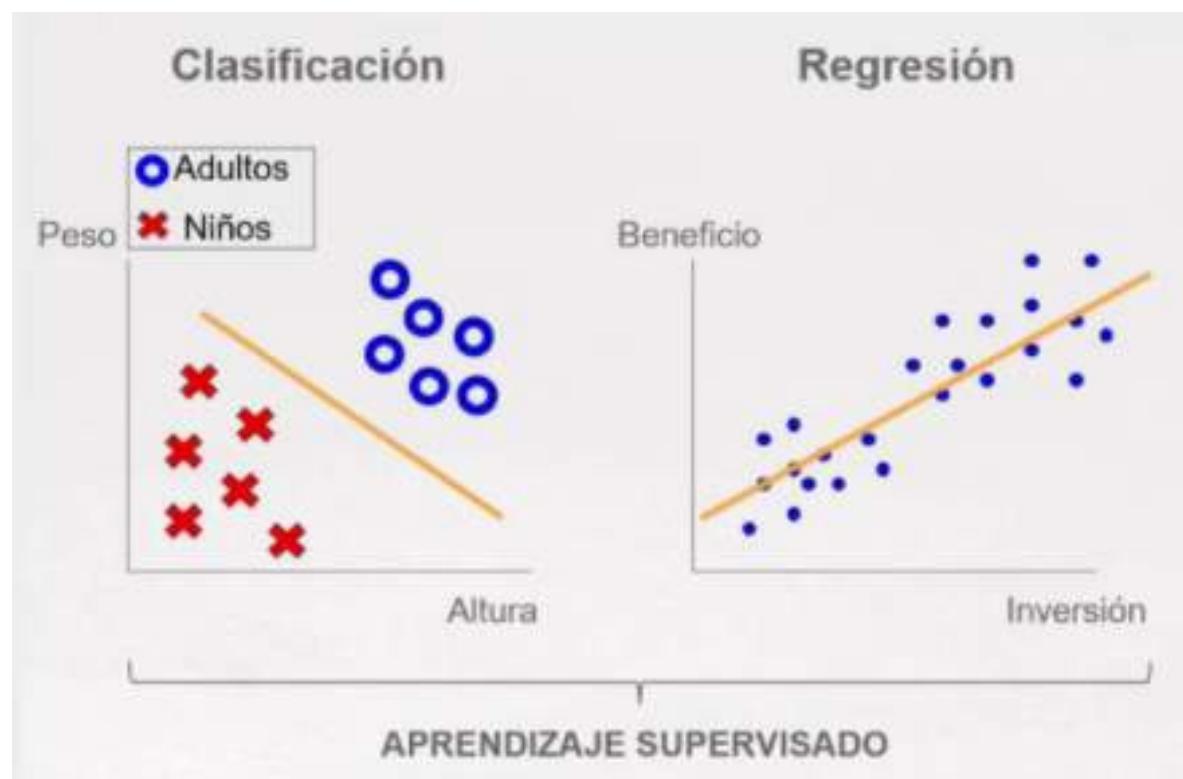
- La **variable objetivo o variable respuesta** es aquella de la que queremos estudiar su comportamiento, bien sea para explicar su variabilidad o para predecir. Suele denotarse como Y, y también se conoce como variable dependiente.
- Las **variables explicativas o independientes** son las que explican el comportamiento de la variable objetivo. Suelen denotarse como  $X_i$ , siendo el subíndice i cada una de variables independientes. En regresión también se las conoce como variables regresoras.

Por ejemplo: queremos explicar las ventas online de una compañía en función de los clics en RR.SS., SEM y display de sus anuncios publicitarios. La variable que queremos explicar (variable objetivo) sería  $Y=\text{ventas de la compañía}$ , y las 3 variables dependientes serían  $X_1=\text{clics RR.SS.}$ ,  $X_2=\text{clics SEM}$ ,  $X_3=\text{clics display}$ .

Los modelos que tienen variable objetivo se engloban dentro del **aprendizaje supervisado** (que, a su vez, se dividen en regresión y clasificación), mientras que aquellos que se centran en las relaciones entre las variables independientes y no existe variable objetivo se engloban dentro del **aprendizaje no supervisado** (subdivididos a su vez en reducción de dimensionalidad y clustering).



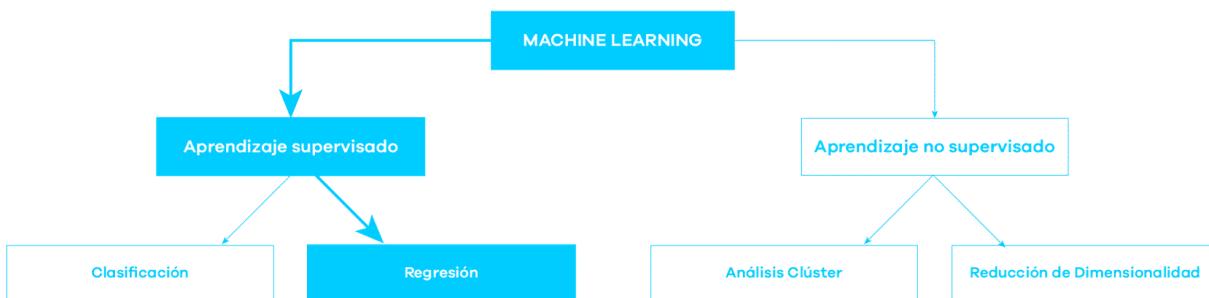
En este fastbook, nos centraremos en los dos modelos básicos de aprendizaje supervisado, la regresión lineal englobada dentro de regresión y la regresión logística englobada dentro de clasificación. En la siguiente imagen verás un ejemplo claro del significado de regresión y clasificación.



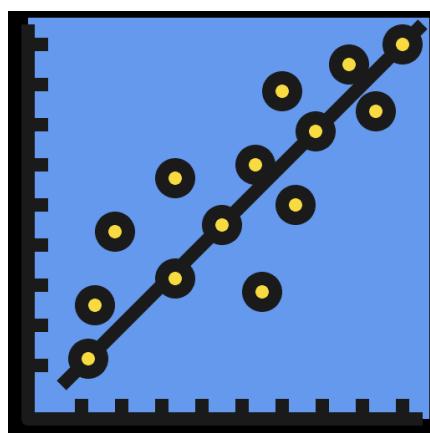
# Regresión lineal

X Edix Educación

Dentro de los modelos de aprendizaje supervisado, el más conocido y utilizado es la regresión lineal. Suele utilizarse como modelo de referencia (benchmark) para evaluar otros modelos más complejos.



A pesar de ser modelos muy sencillos son ampliamente utilizados por sus buenos **resultados predictivos** y la fácil interpretación de las relaciones entre las variables. Se utiliza en prácticamente todas las ramas (economía, ciencias biológicas y de la salud, física, etc.) y sectores (energía, banca, seguros, farma, sanidad, etc.).



La **regresión lineal** tiene buenos resultados cuando queremos predecir, pero realmente destaca por su parte explicativa, es decir, por la explicación del comportamiento de las variables y su relación con la variable objetivo.

La regresión lineal **explica la relación lineal de dependencia** de una variable respuesta Y como combinación lineal de variables explicativas  $X_i$ .

**La regresión lineal modeliza el comportamiento de la variable respuesta mediante una recta.**

A diferencia del contraste de hipótesis (donde se medía la relación de una variable numérica a lo largo de los niveles de una variable categórica), en la regresión lineal la variable respuesta Y y las variables explicativas  $X_j$  son **numéricas y generalmente continuas**.

Dentro de las regresiones lineales podemos distinguir entre **regresión lineal simple** y **regresión lineal múltiple** en función del número de variables independientes  $X_i$  que explican la variable objetivo Y.

# Regresión lineal simple

X Edix Educación

---

La regresión lineal simple es el caso más sencillo de la regresión lineal, donde el comportamiento de la variable respuesta Y es explicada mediante una sola variable independiente X. La formulación matemática es:

$$Y = \beta_0 + \beta_1 X + \text{error}$$

Leyenda de la fórmula.

Y es la variable respuesta o dependiente.

$\beta_0$  se conoce como término independiente, nivel base o intercept. Representa el valor de la variable respuesta Y cuando la variable explicativa X vale 0 ( $X=0 \rightarrow Y=\beta_0$ ).

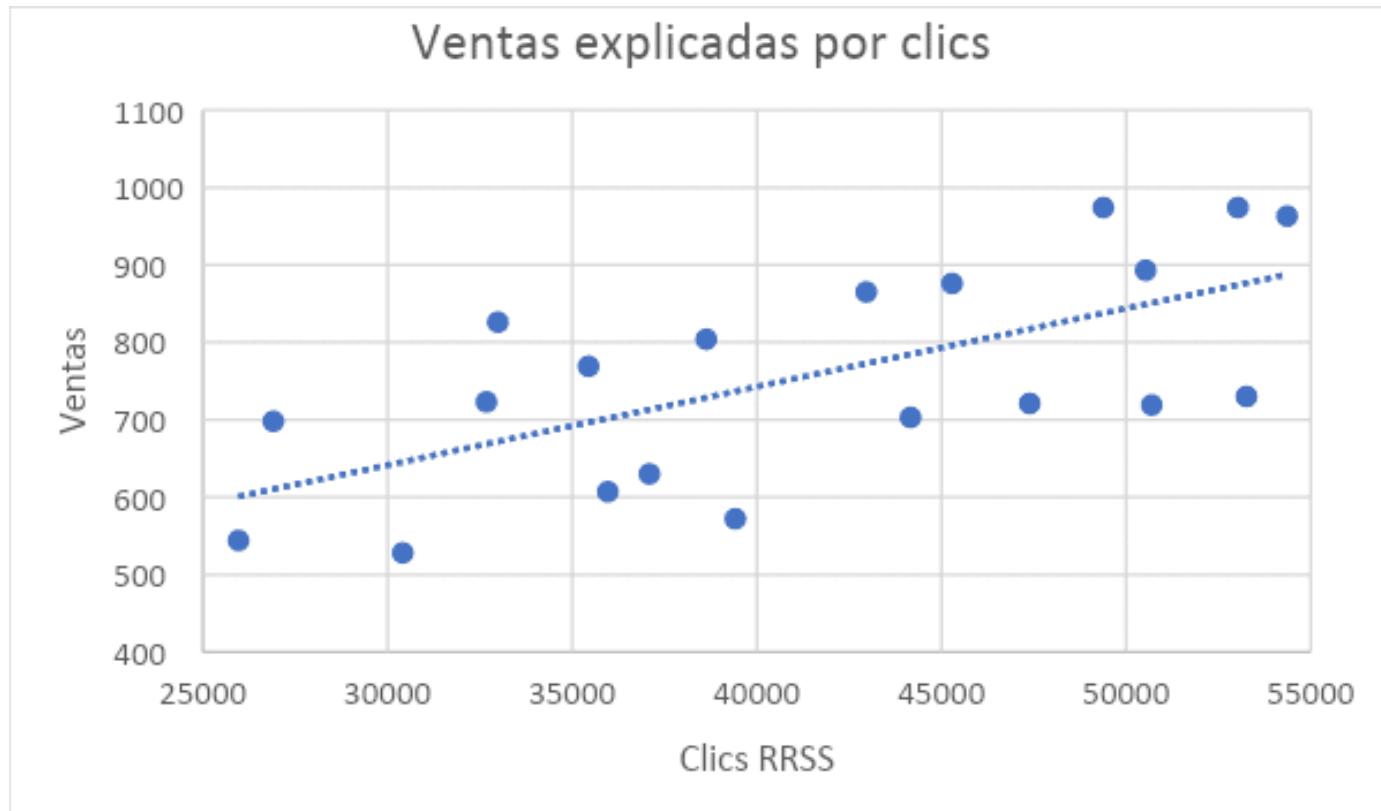
$X$  es la variable explicativa o independiente, también llamada regresora. Con ella explicaremos el comportamiento de la variable respuesta.

$\beta_1$  es la pendiente de la recta de regresión. Por cada unidad que crece  $X$ ,  $Y$  crece  $\beta_1$ .

*error* es el error aleatorio propio de la modelización. No se puede captar toda la influencia de una variable en un modelo por muy elaborado que este sea.

$\beta_0$  y  $\beta_1$  son los parámetros del modelo y se conocen como los **coeficientes de regresión o estimadores**. Miden la importancia o peso de cada variable dependiente sobre la variable objetivo y se estiman minimizando la distancia de los puntos a la recta de regresión. Hay múltiples formas de calcular la recta de regresión, pero entre todas ellas destaca la estimación por mínimos cuadrados ordinarios (MCO) que consiste en minimizar las diferencias al cuadrado de las observaciones a su proyección sobre la recta de regresión.

Veamos un ejemplo. Supongamos que queremos explicar las ventas online de una compañía mediante la publicidad que realiza en redes sociales. Para ello, aplicamos una regresión lineal simple en la que la variable objetivo  $Y=ventas\ online$  es explicada por una variable explicativa  $X=clics\ RRSS$ .



La regresión lineal resultante es  $Y=0,0101X+338,21$ :

- Si no hacemos publicidad en RR.SS. obtenemos unas ventas promedio de 338,21.
- Por cada clic obtendremos 0,01 ventas, o dicho de otra forma, por cada 100 clics obtendremos 1 venta.

Aunque ya sabemos interpretar los coeficientes del modelo, hay más detalles que nos faltan por conocer. Además, para **crear un buen modelo también es necesario medir la bondad de ajuste y la significancia de cada estimador**.

## 1

## Coeficientes

Los **coeficientes del intercept y de las variables explicativas** ( $\beta_0$  y  $\beta_1$ , respectivamente para la regresión lineal simple) miden el peso y el tipo de impacto (positivo o negativo) de cada variable independiente sobre la variable objetivo.

Merece la pena destacar este último aspecto, ya que a veces los coeficientes no están alineados con el impacto real y se obtienen coeficientes negativos cuando tendrían que ser positivos o al revés. Por ejemplo, supongamos que queremos explicar la venta de pisos en función de la tasa de paro. Hacemos una regresión lineal y obtenemos que tanto el  $\beta_0$  como  $\beta_1$  son positivos ( $\beta_0, \beta_1 > 0$ ), pero esto no tiene sentido ya que cuanto más paro hay menos pisos se venden, con lo cual el coeficiente de la tasa de paro debe ser negativo.

## 2

## P-valores

El **p-valor** indica el nivel de significancia de cada variable regresora frente a la variable objetivo. Es un valor que oscila entre 0 y 1 e indica si una variable es significativa a un nivel de confianza  $\alpha$  o no, es decir, si es relevante.

**La lectura del p-valor es la misma que la que veíamos en los contrastes de hipótesis, si el p-valor es más pequeño que  $1-\alpha$  (nivel de confianza) se rechaza la hipótesis nula.**

En este caso el estadístico que suele usarse es la t-student, el nivel de confianza suele ser del 95% y el contraste de hipótesis que se plantea asociado a la pendiente  $\beta_1$  es:

{ $H_0$ : no hay relación lineal entre la variable independiente y la dependiente, por lo que la pendiente es cero:  $1=0$   $H_1$ : sí hay relación lineal entre ambas variables por lo que la pendiente es distinta de cero:  $1\neq 0$ .

En otras palabras: la hipótesis nula indica que no hay relación entre X e Y, mientras que la hipótesis alternativa describe una situación en la que sí hay relación entre ambas variables y, por tanto, el coeficiente asociado es distinto de cero. Por lo tanto, cuando el 1-p-valor  $> \alpha \rightarrow$  aceptamos la hipótesis nula  $H_0 \rightarrow \beta_1=0 \rightarrow$  la variable independiente Y no tiene influencia sobre X.

Cuando planteamos un modelo de regresión lineal, interesa que todas las variables sean significativas, es decir, que se rechace la hipótesis nula y, por tanto, el coeficiente  $\beta_i \neq 0$ .

**Un p-valor menor que 0,05 es lo ideal, aunque estimadores con p-valores  $\leq 0,1$  también son aceptables. Este mismo razonamiento se sigue para  $\beta_0$  y, en el caso multivariante, para el resto de coeficientes  $\beta_j$ .**

### 3

## Ajuste del modelo. Coeficiente de determinación $R^2$

Queremos ahora medir la bondad del modelo, es decir, cuánto de bueno o de malo es. Hay dos medidas importantes que determinan el ajuste de los modelos: la desviación típica residual y el coeficiente de determinación, aunque casi siempre se usa el segundo.

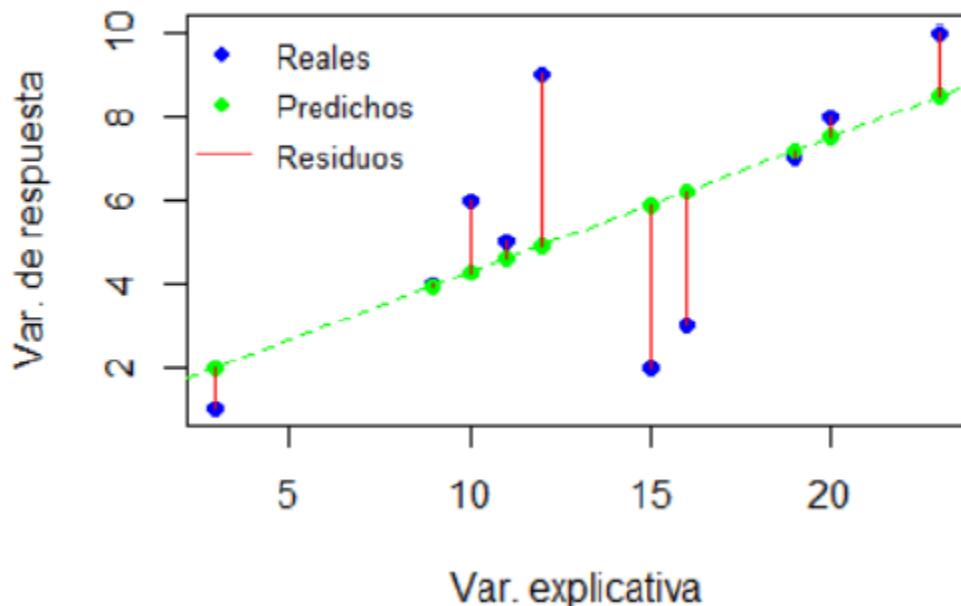
### Desviación típica residual

Representa la variabilidad promedio de los datos observados en relación a la recta de regresión. De forma simple, la distancia de los puntos a la recta de regresión (es la desviación típica que estudiamos en el fastbook 5, pero en lugar de a la media, a la recta de regresión).

## Coeficiente de determinación R<sup>2</sup>

Es el más usado y extendido. Mide lo cerca que está el modelo de la realidad: calculamos las predicciones del modelo aplicando la fórmula de la regresión ( $Y = \beta_0 + \beta_1 X$ ) y lo comparamos con la variable objetivo Y.

El  $R^2$ , como suele llamarse, mide el **porcentaje de variabilidad** de Y explicado por X. Toma valores entre 0 y 1 (o entre 0 y 100): cuanto mayor sea el  $R^2$ , mejor será el ajuste y mayor será la certidumbre que tenemos cuando predecimos Y dando un valor de X, y, por tanto, menor será la desviación típica residual.



$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

### Leyenda de la fórmula:

- $y_i$  son cada uno de los valores que toma la variable respuesta Y (puntos azules).
- $\hat{y}_i$  es el valor de la variable Y dado por el modelo (puntos verdes).
- $\bar{Y}$  es el promedio de  $y_i$ .

La **línea verde es la recta de regresión**, los puntos verdes son los valores que predice el modelo y los puntos azules son los puntos de la variable objetivo. Como vemos, hay **desviaciones entre los puntos azules y verdes** (que son las que determinan el  $R^2$ ) ya que es importante tener en cuenta que el ajuste del modelo a la realidad no es exacto y, por tanto,  $R^2 < 1$ .

En general, suele graficarse la variable objetivo Y y el modelo de regresión para ver el ajuste obtenido, lo que va alineado con el coeficiente de determinación  $R^2$  (ahora verás un ejemplo).

### Ejemplo

Veamos el **ejemplo anterior completo** ahora que tenemos todo el conocimiento necesario: queremos explicar el volumen de ventas online en función de los clics de RR. SS.

Semana	Ventas	Clics RR. SS.
1	528	30.412
2	544	25.965
3	607	35.965

Semana	Ventas	Clcs RR. SS.
4	630	37.091
5	572	39.418
6	698	26.908
7	703	44.164
8	723	32.677
9	719	50.695
10	721	47.392
11	730	53.262
12	769	35.442
13	804	38.631
14	826	32.986
15	974	53.029
16	865	42.966
17	876	45.289

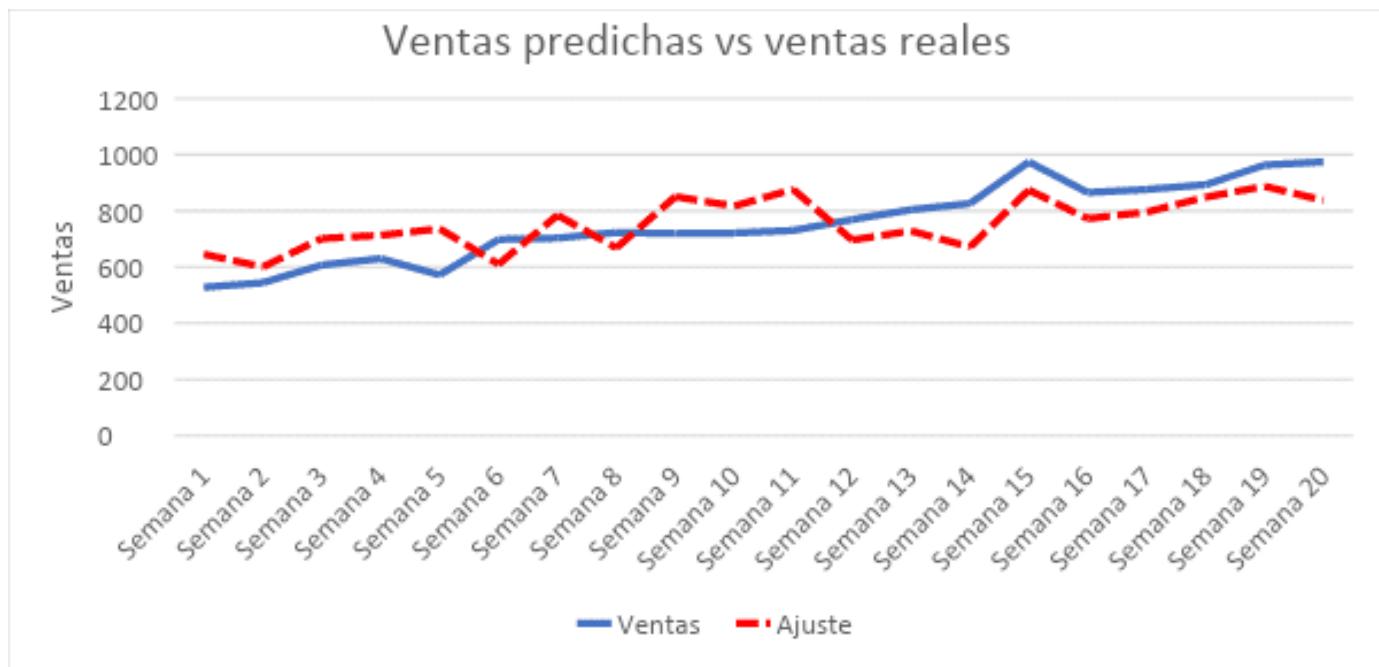
Semana	Ventas	Clcs RR. SS.
18	893	50.531
19	963	54.362
20	974	49.385

Aplicamos una regresión lineal y obtenemos:

SUMMARY OUTPUT				
Regression Statistics				
R Square	0,428			
Observations	20			
	Coefficients	Standard Error	t Stat	P-value
Intercept	338,2130	116,3313	2,9073	0,0094
Clcs RR. SS.	0,0101	0,0028	3,6715	0,0017

Como veíamos anteriormente, los coeficientes (primera columna de la segunda tabla) son  $\beta_0=338,21$  y  $\beta_1=0,0101$ , ambos positivos. Esto tiene sentido porque tanto el nivel base como los clics tienen un impacto positivo sobre las ventas. En la misma tabla, en la segunda columna tenemos el margen de error del estimador, en la tercera el estadístico t, y, en la última, el p-valor, ambos por debajo de 0,05, por lo tanto, ambos son significativos y las variables son relevantes.

De la primera tabla podemos extraer el número de observaciones, en nuestro caso 20. También podemos ver el  $R^2=0,428$ , por lo que, aunque tengamos buenos p-valores de los coeficientes, el ajuste no es bueno. Veámoslo gráficamente:



Como vemos, la línea azul (ventas reales) no tiene un comportamiento parecido a la línea roja discontinua (ventas modelizadas), por lo que el ajuste no es bueno. Podemos concluir que no es un buen modelo ya que no explica gran parte de la variabilidad.

# Regresión lineal múltiple

X Edix Educación

---

Cuando hay más de una variable predictora, la regresión lineal simple se convierte en regresión lineal múltiple: se explica una **variable respuesta Y**, generalmente **continua**, como una **combinación lineal de variables explicativas** ( $X_1, X_2, \dots, X_n$ ) (variables independientes). El **resultado es una línea recta** que se calcula minimizando la distancia de la recta a los puntos correspondientes.

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \text{error}$$

Leyenda de la fórmula.

Y: variable respuesta (dependiente).

$X_1, \dots, X_n$  son las variables explicativas, también llamadas independientes o regresoras.

$\beta_0, \beta_1, \dots, n$  son los coeficientes de las variables explicativas  $X_1, \dots, X_n$ .

error el error aleatorio propio de la modelización.

La regresión lineal múltiple permite analizar de forma simultánea el impacto de varias variables regresoras, lo que, generalmente, permite tener aproximaciones más certeras que la regresión lineal simple.

## 1

### Coeficientes

A diferencia de la regresión lineal simple, el **impacto de cada variable puede no ser directamente comparable** ya que las variables pueden estar en diferentes unidades. Por ejemplo, si queremos modelizar las ventas explicadas por la tasa de paro y el PIB, las unidades de ambas variables son muy diferentes, por lo que cabría esperar que el coeficiente de la tasa de paro sea mucho más alto que el del PIB y el aporte de cada una (medido por el coeficiente) no es comparable.

La solución para este problema es escalar las variables para que ambas tengan el mismo rango y se puedan comparar. Otra alternativa más utilizada, es **medir el impacto de las variables multiplicado por su coeficiente**, es decir, comparar  $\beta_1 X_1$  con  $\beta_2 X_2, \beta_3 X_3, \dots$

2

## P-valor

Al igual que vimos para el caso simple, la significancia de cada variable se mide con un contraste de hipótesis individual para cada coeficiente  $\beta_j$  de cada variable regresora  $X_j$ .

$$\{H_0: \beta_j=0 \quad H_1: \beta_j \neq 0 \quad \forall j=1, \dots, k\}$$

Para que la variable sea significativa, es necesario que el p-valor <  $1-\alpha$  (nivel de confianza, generalmente, de 0,95), y por tanto, el coeficiente de la variable independiente no sea cero:  $\beta_j \neq 0$ .

3

## $R^2$ y $R^2$ ajustado

Para medir la bondad del modelo, en el caso de la regresión lineal múltiple suele usarse el  $R^2$  ajustado. El  $R^2$  aumenta cuando se incluye una nueva variable regresora en el modelo sin importar cuál sea su aportación, por lo que en el caso multivariante se recomienda el uso del coeficiente de determinación ajustado  $R^2_{adj}$  ya que previene modelos sobreajustados.

4

## Multicolinealidad

En ocasiones nos encontraremos con regresores que de forma individual explican la variable dependiente Y.

Estos regresores metidos de forma conjunta en un modelo de regresión lineal múltiple dejan de ser significativos o cambian de signo. Esto es lo que se conoce como multicolinealidad.

Esta situación puede darse debido a que la correlación entre dos regresores es muy alta y lo que trata de explicar uno ya lo ha explicado el otro.

Veamos un ejemplo en el que una variable objetivo se exprese en función de más de una variable independiente. Continuemos con el ejemplo anterior.

Supongamos que queremos explicar las ventas online de una compañía, pero esta vez explicadas por los clics en RR. SS., clics en SEM y clics en display.

Semana	Ventas	Clics RR. SS.	Clics SEM	Clics display
1	528	30.412	49.297	16.034
2	544	25.965	46.034	12.367
3	607	35.965	55.372	17.551
4	630	37.091	47.393	16.965
5	572	39.418	58.288	11.561
6	698	26.908	53.547	9.559
7	703	44.164	61.864	18.085
8	723	32.677	54.143	30.855
9	719	50.695	63.886	20.802
10	721	47.392	68.440	24.346

Semana	Ventas	Clics RR. SS.	Clics SEM	Clics display
11	730	53.262	66.748	26.204
12	769	35.442	58.149	23.286
13	804	38.631	63.361	20.492
14	826	32.986	67.081	28.113
15	974	53.029	70.622	44.220
16	865	42.966	59.398	37.270
17	876	45.289	65.133	30.367
18	893	50.531	73.056	37.110
19	963	54.362	70.817	43.714
20	974	49.385	79.878	40.789

Si aplicamos **regresión lineal multivariante** obtenemos:

SUMMARY OUTPUT				
Regression Statistics				
	Coefficients	Standard Error	t Stat	P-value
R Square	0,878			
Adjusted R Square	0,855			
Standard Error	53,171			
Observations	20			
Intercept	195,7499	97,9335	1,9988	0,0629
Clics RRSS	-0,0029	0,0023	-1,2629	0,2247
Clics SEM	0,0075	0,0025	2,9463	0,0095
Clics display	0,0087	0,0017	5,0420	0,0001

Vemos que el  $R^2_{ajustado}=0.855$ , con lo cual el ajuste es mucho mejor que en la regresión lineal simple (que teníamos un ajuste de 0,428) y los resultados son más fiables. Si estudiamos las variables regresoras, vemos que el coeficiente de los clics de RR.SS. es negativo y su p-valor es ligeramente alto (0,2247). Aunque pudiera parecer un buen modelo, el impacto de los clics de RR.SS. no puede ser negativo sobre las ventas, por lo que tenemos que estudiar la correlación entre las variables al tratarse probablemente de un caso de multicolinealidad.

La matriz de correlación es:

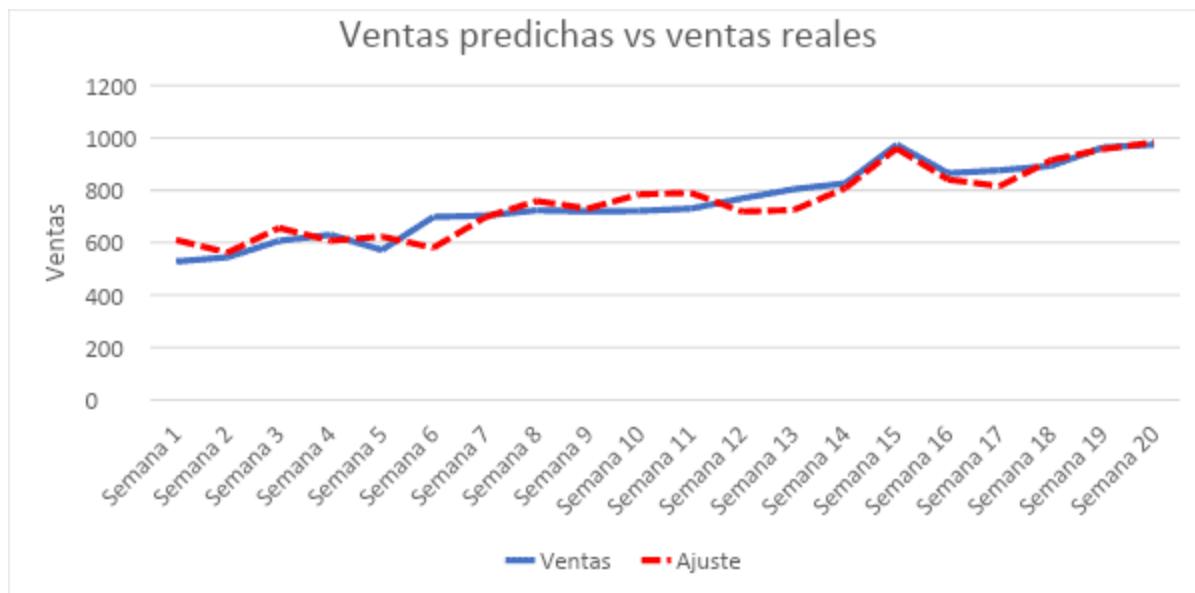
	<i>Ventas</i>	<i>Clics RR. SS.</i>	<i>Clics SEM</i>	<i>Clics display</i>
Ventas	1			
Clics RRSS	0,654	1		
Clics SEM	0,826	0,802	1	
Clics display	0,899	0,682	0,740	1

La correlación de la serie de ventas con todas las variables regresoras es media-alta, pero entre las variables regresoras la correlación también es alta (por ejemplo, 0,802 entre RRSS y SEM). Esto puede ser una explicación de por qué el p-valor de clics RR. SS. es alto y su coeficiente negativo: lo que explica la variable RR.SS. ya lo explican entre las otras dos.

Probamos a sacar esta variable del modelo en búsqueda de un modelo con un buen R2 donde las variables sean significativas y positivas.

SUMMARY OUTPUT				
<i>Regression Statistics</i>				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
R Square	0,865			
Adjusted R Square	0,850			
Standard Error	54,094			
Observations	20			
Intercept	205,9058	99,2956	2,0737	0,0536
Clics SEM	0,0055	0,0021	2,6927	0,0154
Clics display	0,0082	0,0017	4,8005	0,0002

Vemos que el ajuste ha bajado mínimamente (5 milésimas), pero las variables regresoras tienen coeficientes positivos con p-valores bajos, por lo tanto, nos quedaremos con este modelo. Por confirmar, graficaremos el ajuste del modelo para contrastar que está alineado:



Como vemos, el ajuste de nuestro modelo está mucho más cerca de la serie de ventas original.

## 5

### Feature engineering

El **feature engineering** consiste en la transformación de las variables originales en otras más sofisticadas con la finalidad de mejorar el ajuste del modelo y las relaciones entre las variables (también evita problemas de multicolinealidad). De esta forma, el modelo sigue siendo lineal, aunque las variables estén transformadas.

Para aplicar la transformación correcta, un buen análisis descriptivo junto con el conocimiento de negocio y comportamiento de cada variable son claves para identificar la transformación adecuada.

Las más usuales son:

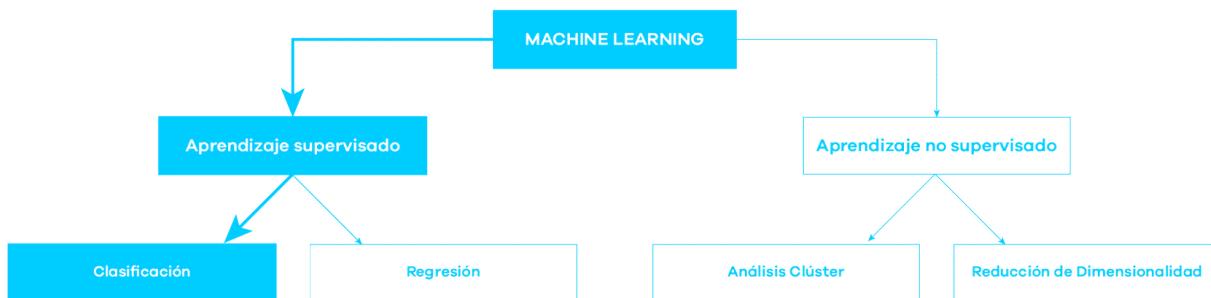
- **Función logarítmica** para evitar valores extremos y suavizar el comportamiento.
- **Centrar los datos** (sustituimos  $x_i$  por  $x_i - \bar{x}_i$ ) para reducir colinealidad.
- **Polinomios** de cualquier grado, aunque usualmente se usan de grado 2.
- **Transformar** una variable categórica en dummy.
- Crear una **nueva variable como multiplicación** de una dummy por una continua.
- Transformaciones a la **variable objetivo**, como, por ejemplo, el logaritmo, la inversa, centrar, etc.

En cualquiera de las transformaciones, una vez ajustado el modelo hay que deshacer el cambio para volver a tener las variables originales.

# Regresión logística

X Edix Educación

Una vez comprendido el ciclo completo de la regresión lineal, pasemos a la regresión logística. Recordemos que se ubicaba dentro de los algoritmos de aprendizaje supervisado, dentro de clasificación.



Los modelos de regresión logística son probablemente los más extendidos dentro del aprendizaje supervisado de clasificación. Su finalidad es determinar el estado de la variable objetivo entre dos opciones. En la regresión logística la variable dependiente es binaria, es decir, toma solamente 2 valores, mientras que las variables independientes pueden ser numéricas o discretas.

Algunos de los **ejemplos más comunes** son la predicción de fuga, la propensión a compra, la probabilidad de impago, etc. En todos, la variable objetivo puede tomar dos valores:

$Y=\{0 \text{ o } 1\}$  (no se va de la empresa, no compra, paga, etc.) 0 o sí (se va de la empresa, compra, hace impago, etc.).

Modelizar una variable que pueda tomar solamente 2 valores es complicado, por lo que modelizaremos la probabilidad de que suceda uno de los eventos:

$$P(X=x) = p(x) \quad P(X=\bar{x}) = 1 - p(x)$$

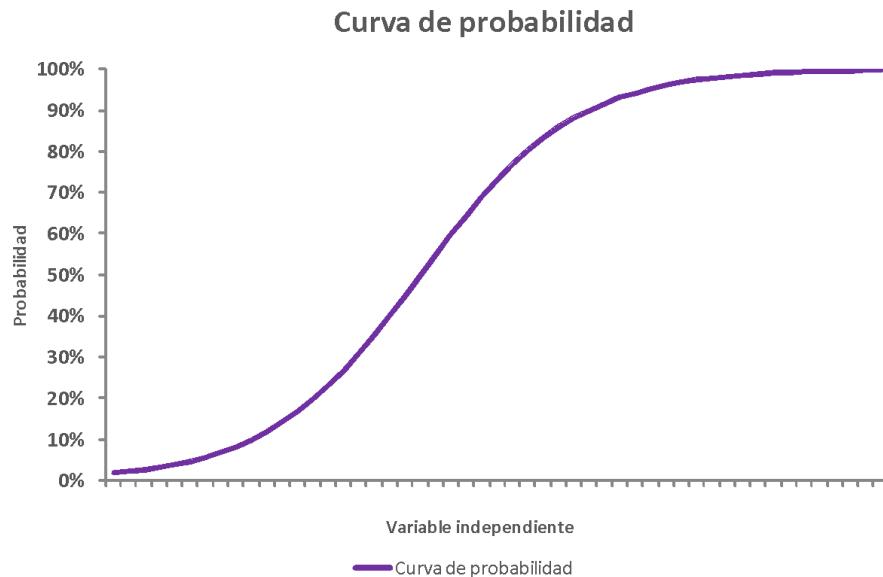
En este caso, estamos modelizando una probabilidad, por lo que  $0 \leq p(x) \leq 1$ .  $\bar{x}$

Si ahora aplicamos una transformación básica conocida como odd,  $p/(1-p)$ , y a esta le aplicamos el logaritmo  $\log(p/(1-p))$ , podemos modelizar nuestra variable objetivo como si fuera una regresión lineal ya que estaremos modelizando una variable numérica continua:  
 $-\infty \leq \log \log(p/(1-p)) \leq +\infty$ .

Tras estas sencillas transformaciones, hemos obtenido la fórmula de la regresión logística basada en la regresión lineal:

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}}$$

Una regresión logística gráficamente tiene la siguiente forma:



En resumen: con una **transformación basada en el logaritmo de los odd** hemos conseguido modelizar una variable categórica con la simplicidad de un modelo de regresión lineal, y, por tanto, el comportamiento en términos de coeficientes y p-valores es igual que para la regresión lineal.

Respecto a la lectura de los resultados, no es tan trivial como en el caso de la regresión lineal. Hay varias formas de leer los resultados de la regresión logística en función de si hablamos de  $p(x)$ ,  $p(x)/(1-p(x))$  o  $\log \log(p(x)/(1-p(x)))$ . Aunque pueda parecer raro, una vez entendida es igual de sencilla.

**La manera más común y fácil de entender es la basada en los odd-ratio y mide el incremento o decremento de probabilidad de que suceda el evento en función de la variable independiente.**

Es decir, si el odd ratio de la variable ‘es menor de 30 años’ es 3, significa que cuando la observación tenga la variable ‘es menor de 30 años’=1 la probabilidad del evento es 3 veces mayor que si no lo es. Más adelante veremos un ejemplo con el que lo entenderás mejor.

Para pasar de los coeficientes que da el modelo a los odd-ratio, simplemente hay que aplicar la exponencial. Con esta transformación, aquellas variables que tenían un coeficiente negativo, su odd-ratio será menor que 1, lo que significará que la probabilidad de que el evento suceda es menor que en el resto de casos. Aquellos coeficientes positivos tendrán un odd-ratio superior a 1, lo que significa que la probabilidad de que suceda el evento es mayor que si no lo tiene.

## Ajuste del modelo

La diferencia respecto a los modelos de regresión lineal reside en las métricas de ajuste de los modelos, ya que estamos ante un problema de clasificación y no de regresión.

Hay **múltiples formas de medir la bondad de los modelos**, pero entre todas ellas destacar el accuracy y la curva ROC junto con el AUC.

1

Matriz de confusión: accuracy

La **matriz de confusión** es de gran utilidad para cualquier problema de clasificación. Se trata de crear una matriz con las frecuencias (o los porcentajes asociados) para entender el comportamiento predicción-realidad y así comprender el ajuste del modelo y los principales problemas. Tiene la siguiente forma:

		Predicción	
		0 (predicción negativa)	1 (predicción positiva)
Realidad	0 (realidad negativa)	TN (verdaderos negativos)	FP (falsos positivos)
	1 (realidad positiva)	FN (falsos negativos)	TP (verdaderos positivos)

**TN (true positive)**

Valores bien clasificados, son negativos y se predicen/clasifican como negativos.

**TP (true positive)**

Valores bien clasificados, son positivos y se predicen/clasifican como positivos.

### **FP (false positive)**

Valores bien clasificados, son negativos y se predicen/clasifican como positivos.

### **FN (false negativo)**

Valores bien clasificados, son positivos y se predicen/clasifican como negativos.

Como se puede ver, es una forma fácil y rápida de saber cuánto de bueno o de malo es el modelo según la ubicación de las observaciones: cuántas más haya en TN y TP mejor, ya que son valores bien clasificados.

Una vez entendamos la **matriz de confusión**, se pueden crear muchas métricas a partir de ella: sensibilidad o recall, especificidad, precisión, valor F1 o accuracy. De entre todas ellas, centraremos la atención en el accuracy.

---

El **accuracy** es una métrica intuitiva y fácil de entender que mide el porcentaje de observaciones bien clasificadas, tanto positivas como negativas.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

La principal desventaja es que suele ser muy engañosa cuando la proporción de observaciones de cada categoría no es similar (muestra desbalanceada).

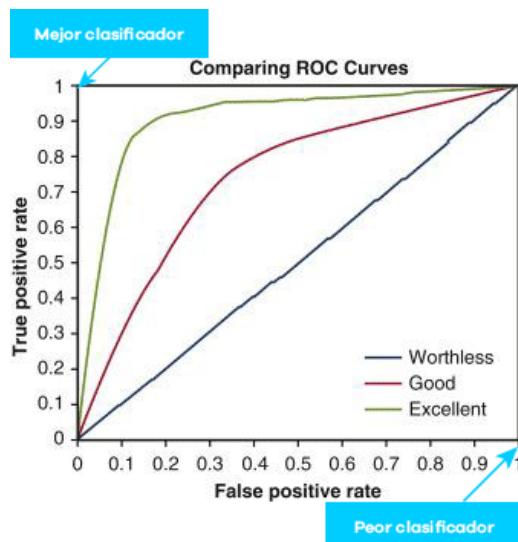
## 2

## Curva ROC y AUC

La **curva ROC** es una herramienta gráfica que permite de un vistazo conocer la bondad del modelo. Representa la proporción de verdaderos positivos (TPR) frente a la proporción de falsos positivos (FPR) para diferentes puntos de corte. Los puntos de corte son ‘probabilidades’ de corte para determinar si una observación es clasificada como positiva o negativa.

$$TPR = \text{sensibilidad} = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{FP+TN} = 1 - \text{especificidad}$$



Asociado a la **curva ROC** tenemos el **área debajo de la curva (AUC: area under the curve)** que mide de forma numérica la bondad del modelo para discriminar observaciones positivas y negativas a lo largo de todo el rango de puntos de corte posibles. Es un indicador que toma valores entre 0 y 1.

3

AIC

El **criterio de información de Akaike (AIC)** es una medida de la calidad de un modelo estadístico que ayuda a decidir cuál es el mejor modelo.

El AIC es un indicador que da un valor numérico basado en la bondad del ajuste del modelo y su complejidad. Cuanto menor sea este valor, mejor es el modelo.

Como curiosidad, la fórmula matemática está basada en el número de parámetros k y el máximo de la función de verosimilitud L.

$$AIC=2K-2\ln(L)$$

Supongamos que somos del departamento de RR.HH. de una gran empresa de carácter internacional y tenemos algunas vacantes abiertas. Queremos determinar cuáles son las variables que más impactan en la contratación de los candidatos. Para ello, aplicamos una regresión logística obteniendo los siguientes resultados:

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
Intercept	-1.64484	0.05622	-29.255	< 2e-16 ***
Nota promedio superior a 7	0.05952	0.03885	1.532	0.12547
Menor de 30 años	0.59716	0.03357	17.787	< 2e-16 ***
Mayor de 55 años	-0.55295	0.03224	-17.148	< 2e-16 ***
Vive a más de 1h de la oficina.	-0.12203	0.04070	-2.998	0.00272 **
Mayor de 45 años	-0.07515	0.03330	-2.257	0.02403 *
Conoce la empresa	0.39553	0.06688	5.914	3.34e-09 ***
Tiene conocimientos informáticos	0.30396	0.05792	5.248	1.54e-07 ***
No quiere movilidad internacional	-0.05044	0.03218	-1.568	0.11697
Habla inglés	0.97962	0.03893	25.163	< 2e-16 ***
Tiene conocimientos de matemáticas	0.27888	0.03043	9.166	< 2e-16 ***
Puesto para Francia	-0.21173	0.05105	-4.148	3.36e-05 ***
Puesto para España	-0.31713	0.04819	-6.580	4.69e-11 ***
Puesto para UK	-0.53069	0.08451	-6.279	3.40e-10 ***

--  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 39125 on 37362 degrees of freedom  
 Residual deviance: 35461 on 37349 degrees of freedom  
 AIC: 35489

Los coeficientes van alineados con el conocimiento que tenemos de la empresa: la lejanía, la edad y la ubicación del puesto para el que aplica (en estos países la contratación es más complicada ya que son sedes con alta demanda) son factores que pueden dificultar la contratación y por tanto, tienen coeficientes negativos. Por otro lado, el resto de variables tienen coeficiente positivo porque favorecen la contratación de la persona.

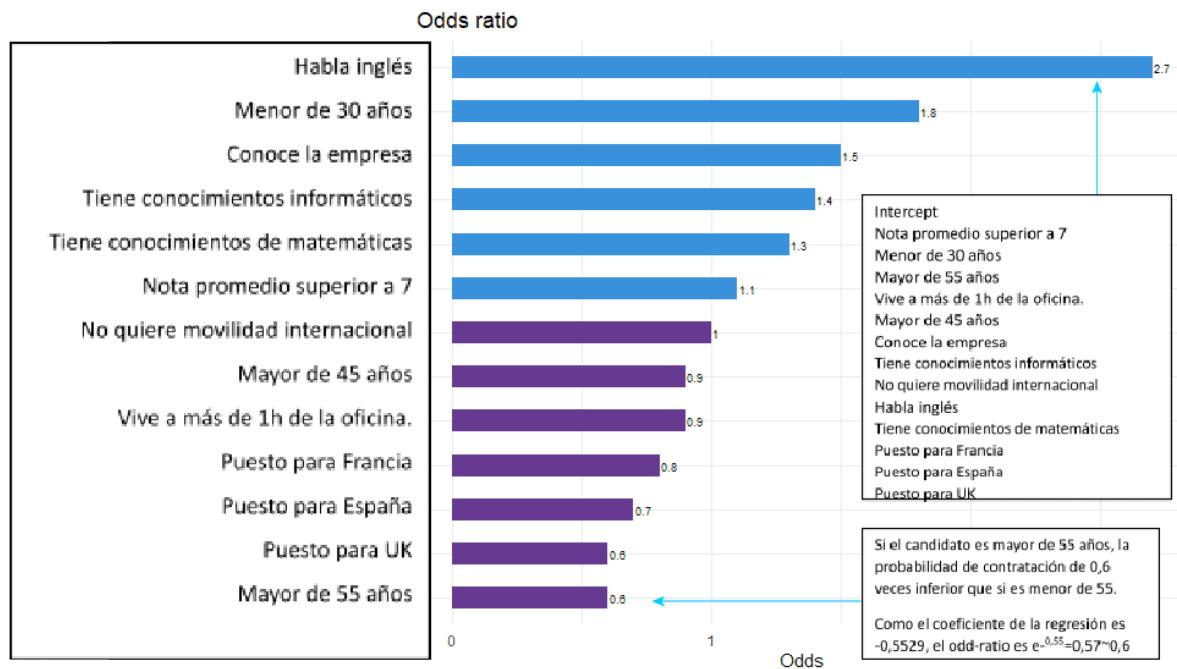
Como vemos, casi todas las variables son significativas, excepto ‘conoce la empresa’ y ‘no quiere movilidad internacional’, pero tienen p-valores aceptables inferiores a 0,2.

Aquí podemos ver el AIC, pero para que tenga sentido necesitamos más modelos para saber si es bueno o malo.

Como hemos dicho anteriormente, la mejor forma de leer el impacto de las variables es mediante los odd-ratio. Recordemos que al aplicar la transformación de los odds, los

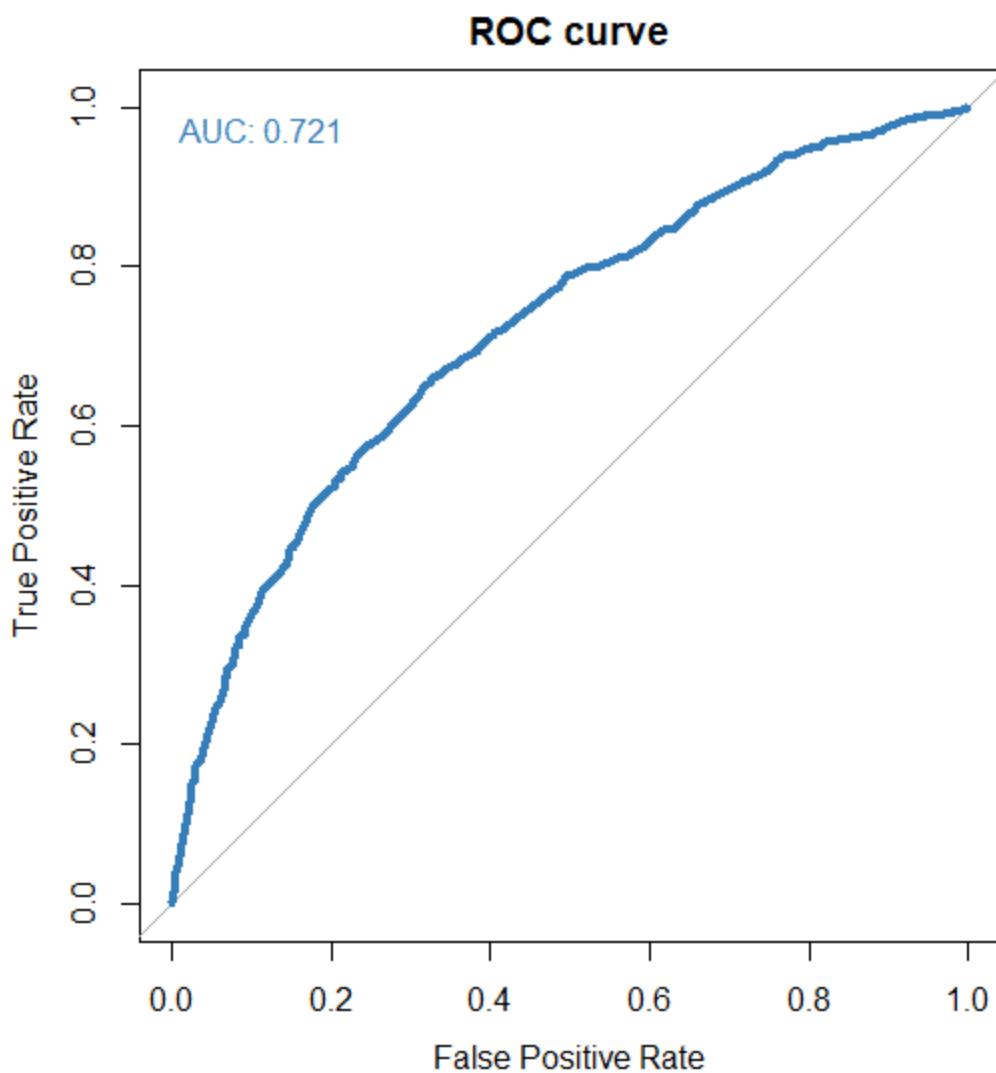
coeficientes negativos pasará a tener valores menores que 1 y los coeficientes positivos pasarán a ser mayores que 1.

Si graficamos los **odd-ratio** (recordemos que se calculan aplicando la exponencial a los coeficientes) veremos de un vistazo el impacto de cada variable. Las variables con barras en verde tienen impacto positivo (coeficientes > 0), mientras que aquellas con barras en rojo tienen impacto negativo (coeficientes < 0).



Como vemos, la variable que mayor impacto positivo tiene es saber hablar inglés seguido de ser menor de 30 años y conocer la empresa. Por otro lado, los perfiles mayores de 55 años o que aplican a puestos para UK tienen una probabilidad menor de ser contratados.

Por último, estudiamos la curva ROC y el AUC:



Como vemos, el modelo tiene un AUC de 0,721, por lo que tiene un buen ajuste.

# Resumen

X Edix Educación

---

En este fastbook hemos comprendido la utilidad de los modelos de regresión lineal y de regresión logística:

- En la **regresión lineal** hemos comprendido la importancia de los coeficientes, cómo leer los p-valores, y cómo medir el ajuste de los modelos.
- En la **regresión logística** hemos aprendido su utilidad y forma de calcularlo, el significado de odd-ratio, cómo medir el ajuste y cómo leer los modelos.

Un resumen de las definiciones mostradas a lo largo de este fastbook:

Regresión lineal	Modelo de regresión que consiste en la combinación lineal de variables continuas para modelizar o predecir una variable objetivo numérica continua.
Coeficientes	Valor a estimar que mide el impacto de cada variable independiente sobre la variable objetivo.
p-valor	Indicador de la relevancia de una variable.
R <sup>2</sup>	Medida de ajuste adecuada para la regresión lineal simple.

<b>R<sup>2</sup> ajustado</b>	Medida de ajuste enfocada para la regresión lineal múltiple.
<b>Multicolinealidad</b>	Efecto que se produce entre dos o más variables regresoras que tienen dependencia entre ellas y provoca un p-valor alto o un coeficiente negativo.
<b>Feature engineering</b>	Transformaciones de las variables originales para mejorar el ajuste del modelo y las relaciones entre variables.
<b>Regresión logística</b>	Modelo de clasificación que consiste en la combinación lineal de variables continuas para modelizar o predecir la probabilidad de que suceda un evento (variable respuesta binaria).
<b>Odd-ratio</b>	Transformación de la variable respuesta en regresión logística: $p/(1-p)$ .
<b>AUC</b>	Métrica asociada a la curva ROC que mide el ajuste del modelo.
<b>Accuracy</b>	Métrica que mide el porcentaje de observaciones bien clasificadas.
<b>Matriz de confusión</b>	Matriz de frecuencias o porcentajes en problemas de clasificación que representa el comportamiento predicción-realidad.

¡Enhorabuena! Fastbook superado

edix

Creamos Digital Workers