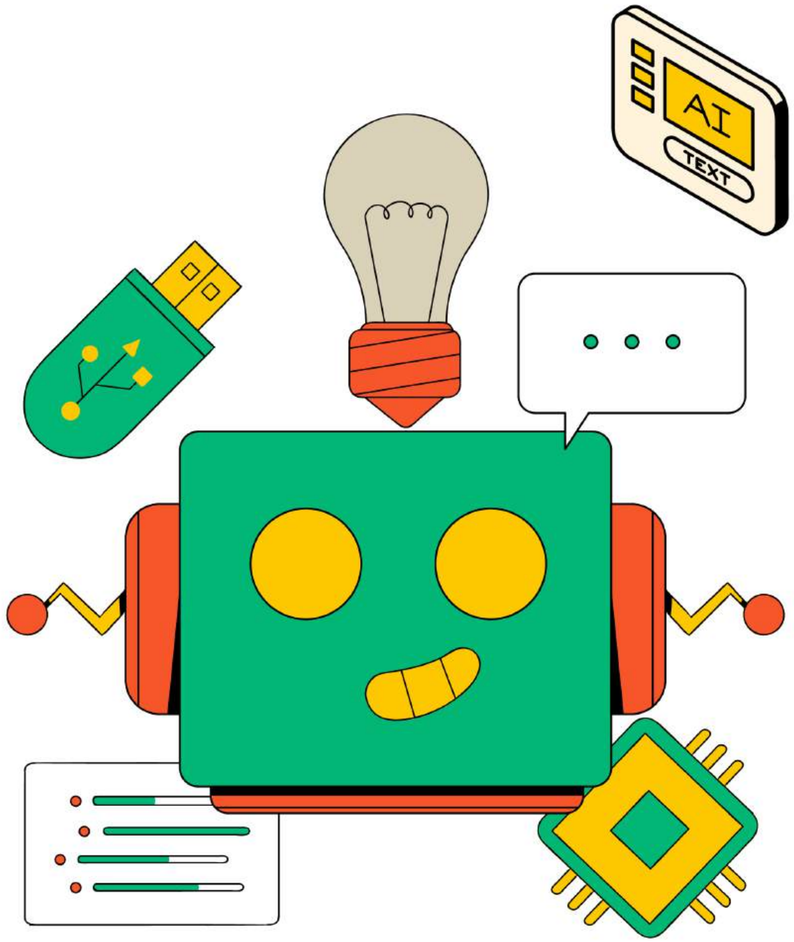
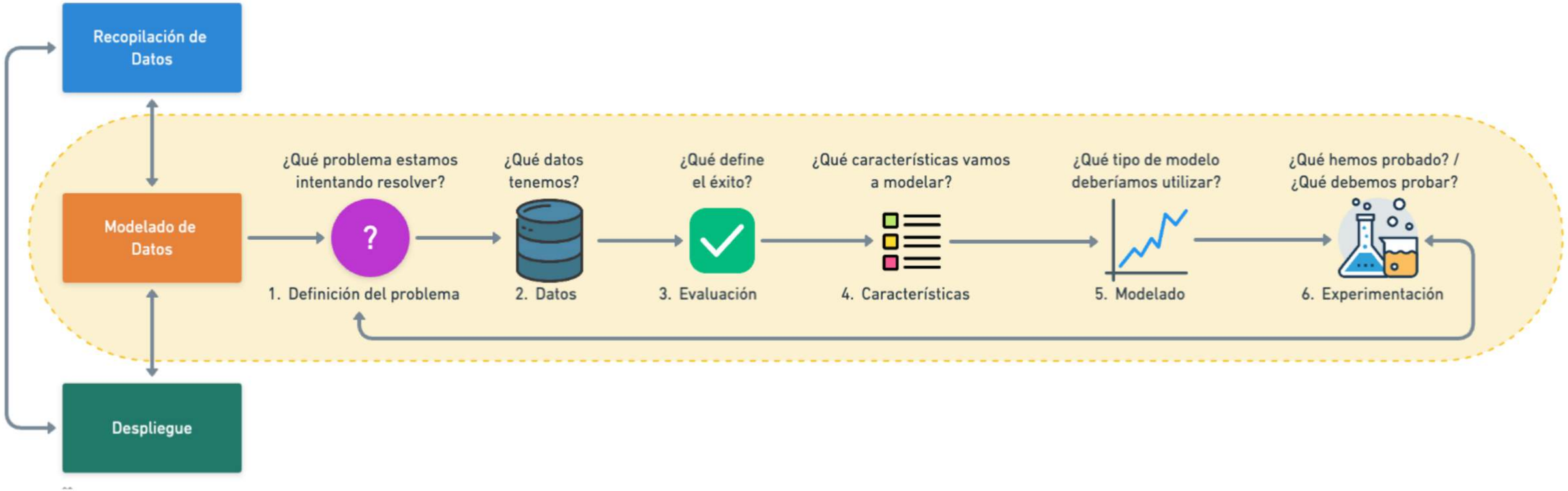
IBERO
CIUDAD DE MÉXICO ®



MACHINE LEARNING

MTRO. ALFONSO GREGORIO RIVERO DUARTE

FRAMEWORK


Recopilación de Datos
Modelado de Datos
Despliegue
¿Qué problema estamos intentando resolver?
1. Definición del problema
¿Qué datos tenemos?
2. Datos
¿Qué define el éxito?
3. Evaluación
¿Qué características vamos a modelar?
4. Características
¿Qué tipo de modelo deberíamos utilizar?
5. Modelado
¿Qué hemos probado? / ¿Qué debemos probar?
6. Experimentación

MODELOS DE CLASIFICACIÓN

A diferencia de la regresión donde se predice un valor continuo, se utiliza la clasificación para predecir una categoría. Existen una gran amalgama de aplicaciones del proceso de clasificación desde medicina hasta marketing. Los modelos de clasificación incluyen desde modelos lineales como la Regresión Logística, SVM, así como otros no lineales como K-NN, Kernel SVM y Bosques Aleatorios.

TÉCNICAS O ALGORITMOS

MODELOS DE REGRESIÓN

1. Regresión Logística
2. K-Nearest Neighbors (K-NN)
3. Support Vector Machine (SVM)
4. Kernel SVM
5. Naive Bayes
6. Árboles de Decisión para Clasificación
7. Clasificación con Bosques Aleatorios

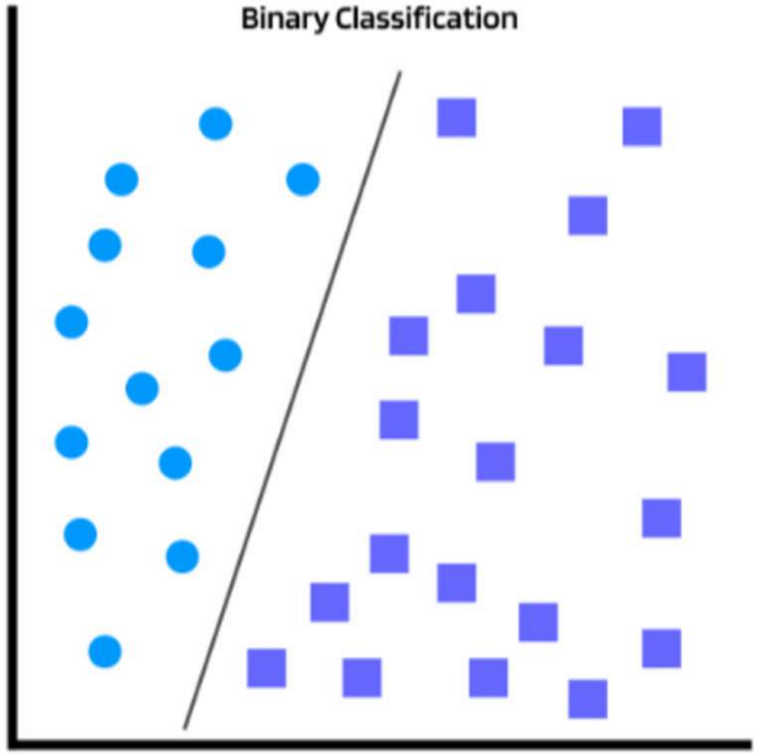
REGRESIÓN LOGÍSTICA O CLASIFICACIÓN

MODELOS DE REGRESIÓN

1. Idea de la Regresión Logística
2. Descripción del problema
3. Algoritmo de Regresión Logística
4. Ajuste del modelo
5. Evaluación

IDEA DE LA REGRESIÓN LOGÍSTICA O CLASIFICACIÓN

CLASIFICACIÓN

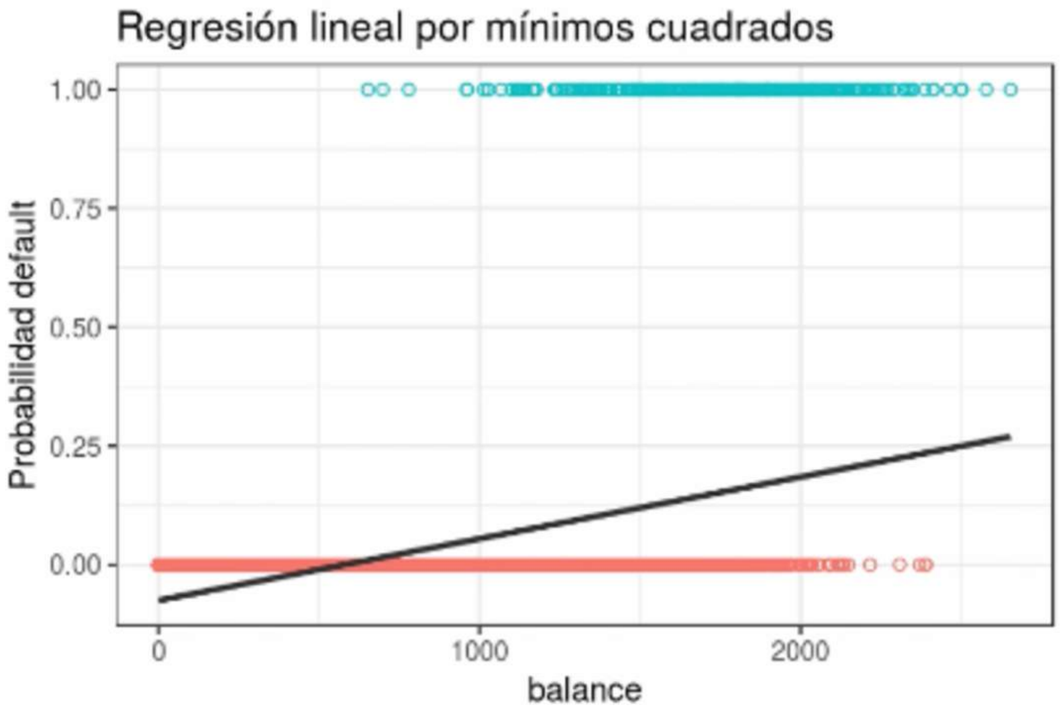


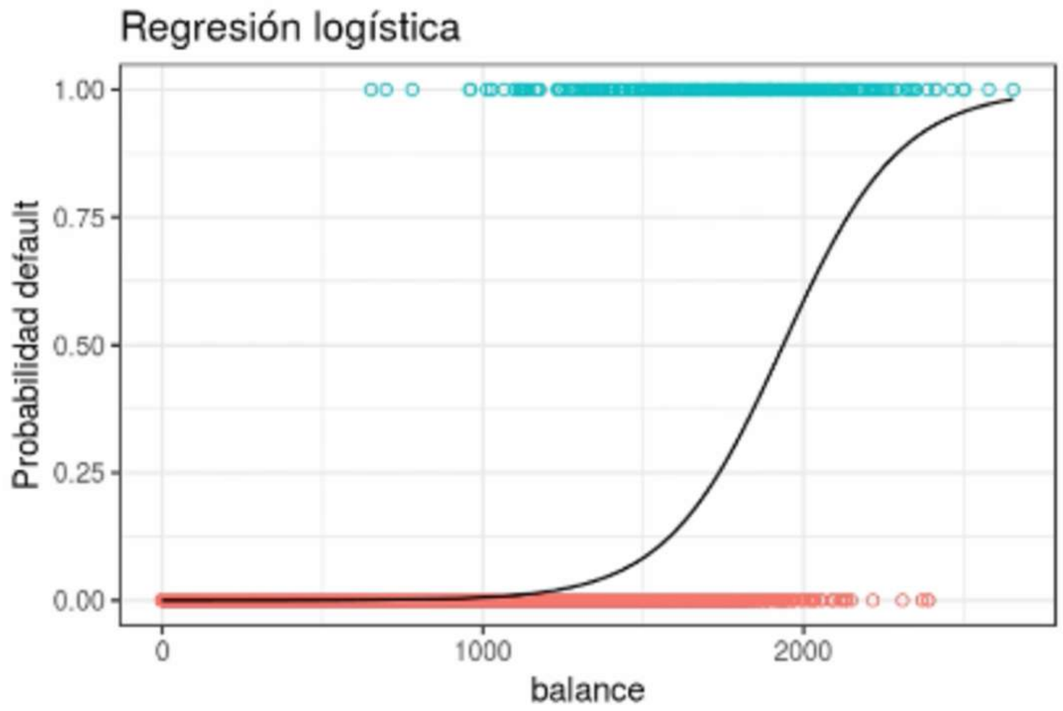
Es un modelo de regresión pero para estimar la probabilidad de un resultario binario:

- Diagnosis de enfermedades
- Correo spam o no
- Si o no para la devolución de un préstamo

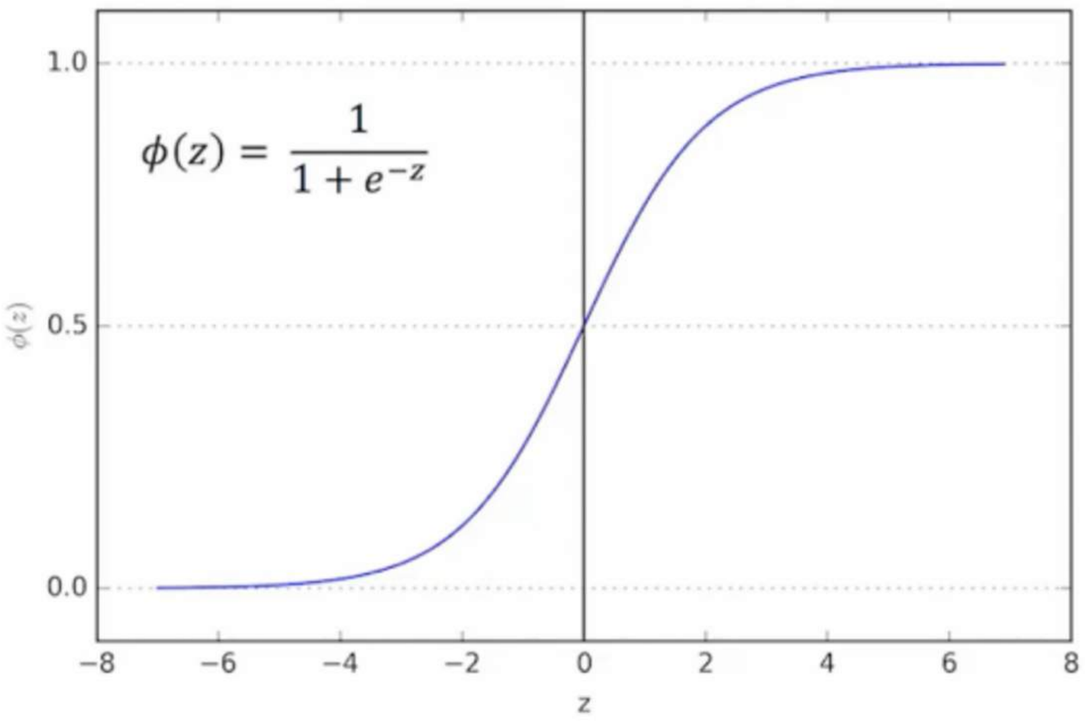
Te permite jugar con las métricas para mejor o peor Sensibilidad o Especifidad

CLASIFICACIÓN





CLASIFICACIÓN



Aprendizaje supervisado

Aprendizaje basado en modelos

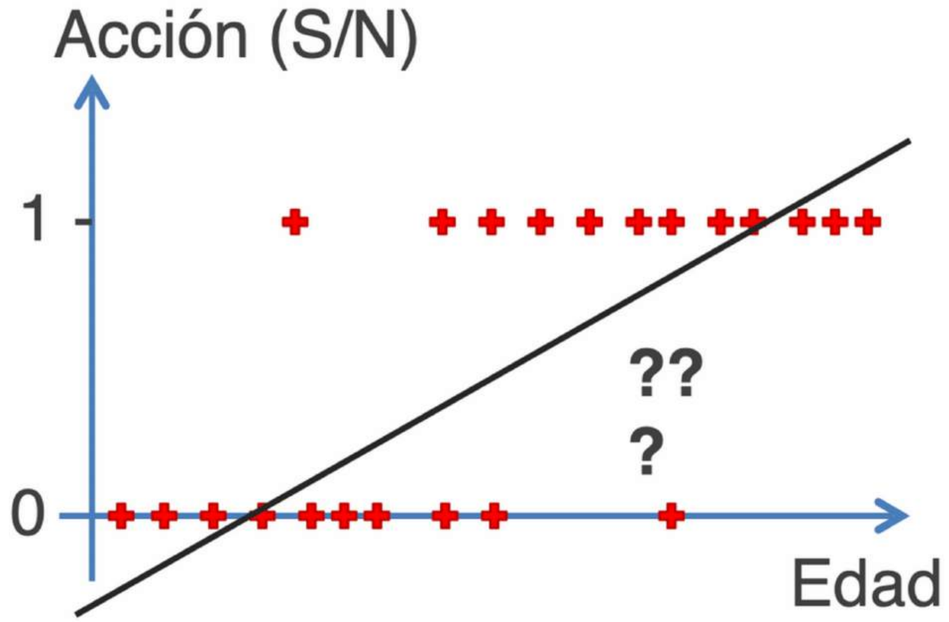
Se corresponde con un modelo lineal generalizado

Realiza predicciones computando una suma ponderada de las características de entrada y sumandole una constante conocida como bias, pero se aplica una función logística al resutlado

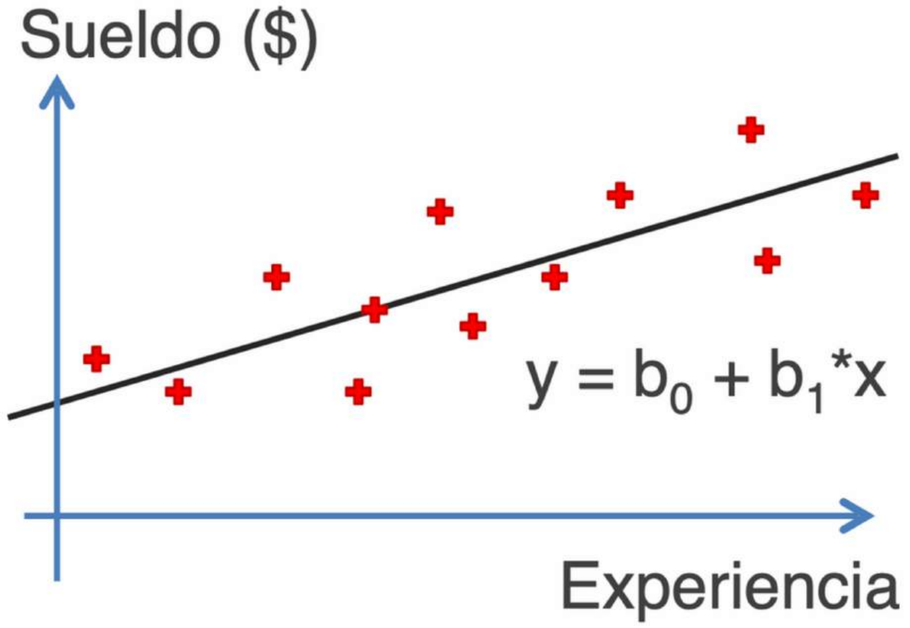
Intenta predecir valores discretos

CLASIFICACIÓN

Lo nuevo es:



Sabemos que:



Imaginemos que ahora enviamos ofertas por correo y esperamos una respuesta SI o NO por parte del cliente.

Tenemos la respuesta de si compro o no utilizando como variable independiente la edad.

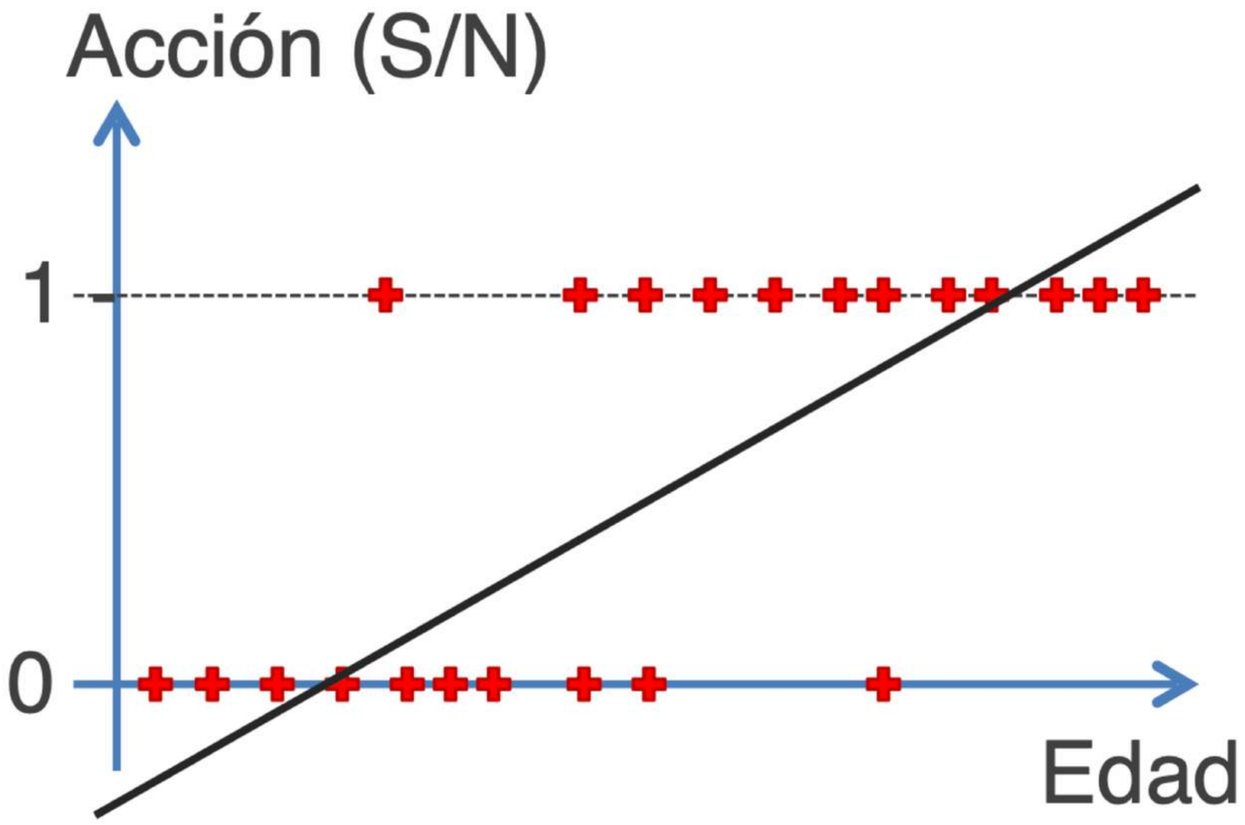
CLASIFICACIÓN

Entonces la pregunta es lleva a cabo una acción el comprar, no comprar el producto en función de la edad?

Al mismo tiempo, aquí de momento, con los conocimientos que tenemos, no sabemos cómo enfocar este problema, no sabemos lo que estamos pronosticando, incluso no podemos ver una correlación numérica en el sentido que conocemos.

En la parte superior está más tirando a la derecha, lo que indica probablemente que las personas mayores tengan más probabilidades de llevar a cabo la acción o tomar la oferta, y las personas más jóvenes son más propensas a ignorarlas.

CLASIFICACIÓN



CLASIFICACIÓN

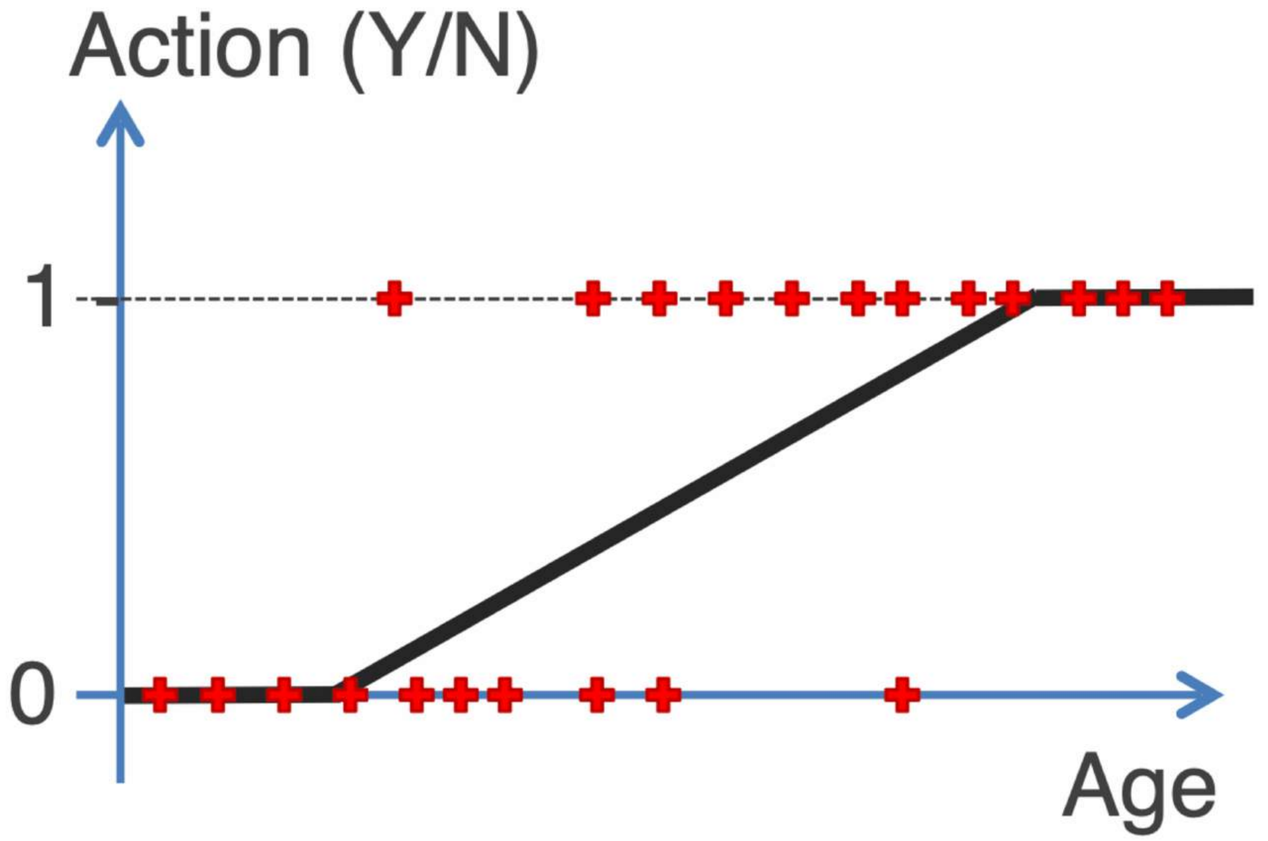
Entonces, una regresión lineal pasaría por trazar una recta. Esto sería la recta de regresión, pero yo creo que lo que obtenemos no sería el mejor método para resolver este problema.

Si nos enfocamos en el problema, pues fijaros, yo podría dibujar una línea horizontal exactamente en el valor que representa la compra, pero el problema sería de predecir exactamente qué va a ocurrir con una persona que tenga una cierta edad de si va a llevar a cabo o no la acción.

En este caso la recta de regresión se podría tunear, se podría afinar algo de este estilo, de si realmente afecta, aceptará o no la oferta truncando la recta de regresión.

Entonces, en lugar de predecir exactamente lo que va a suceder, podríamos intentar predecir las reglas de probabilidad.

CLASIFICACIÓN



CLASIFICACIÓN

¿Qué tan probable es que un cliente acepte la oferta?

Estos tres trocitos de r cta empiezan a acercarse y empiezan a descubrir un poquito de luz de cuál va a ser el enfoque de que el resultado no tiene que ser una recta de regresión, sino que, como las probabilidades van a estar entre cero y uno, todo lo que esté por debajo del cero será transformado a un valor cero que se traduce en no compra. El valor uno será traducido automáticamente en si compra.

Lo que básicamente nos está diciendo esta especie de recta de regresión es que en función de la edad, es o no es probable que ocurra la compra.

La idea es encontrar esas franjas de edades y cuantificar esas probabilidades de aceptar o no la oferta.

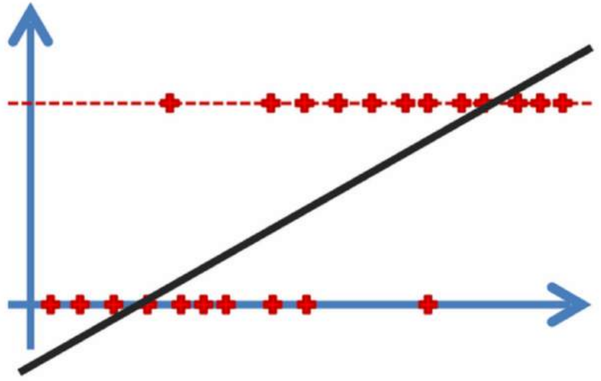
CLASIFICACIÓN

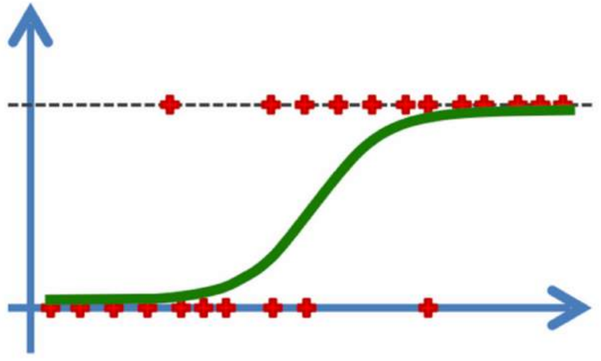
$$y = b_0 + b_1 * x$$

Función Sigmoide

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$





En lugar de una regresión lineal, se aplica una función sigmoide

CLASIFICACIÓN

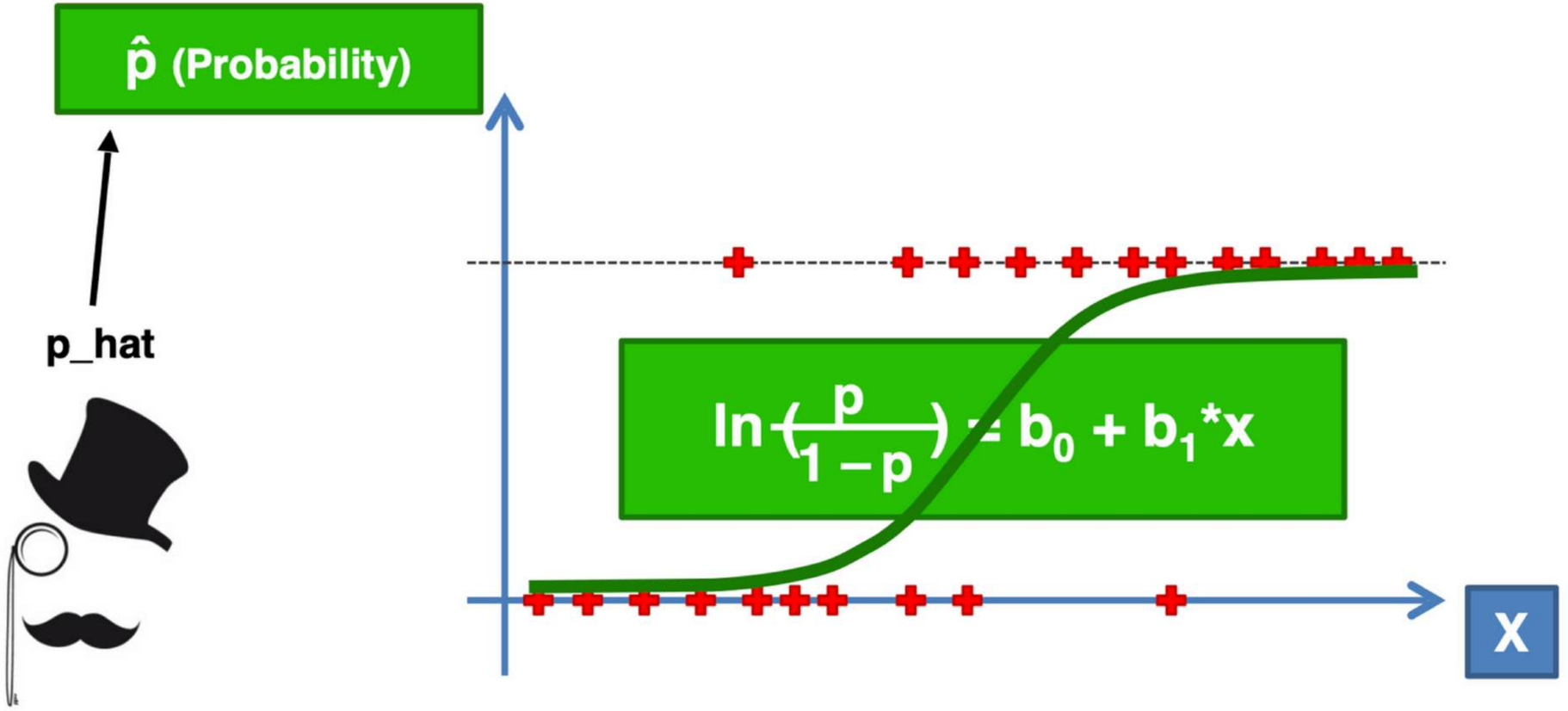
Entonces tenemos expresada una combinación lineal, una regresión lineal, pero a la predicción se le aplica una función sigmoide para transformar el valor final de la regresión en una probabilidad.

De ahi viene el nombre de regresión logística.

La variable dependiente tomaría una serie de valores en función de las observaciones que tengamos en el conjunto de datos. Aquí representamos las observaciones de compra en función de la edad.

Aparece esta curva, este especie de tuneo a la regresión lineal. Y básicamente es una función que se interpreta del mismo modo que la pendiente o la línea de tendencia de una regresión lineal, pero con la diferencia de que está un poquito curva y lo que hace es modelar la tendencia, en este caso de compra o no compra.

CLASIFICACIÓN



CLASIFICACIÓN

Básicamente hacemos lo mismo que la recta de regresión lineal, pero se ve diferente porque es un proceso de clasificación. De las muchas curvas de este estilo que empezaran abajo, acabarán arriba, y se tuercen en la zona de en medio, buscamos cuál es la que mejor se ajusta a los puntos que tenemos y eso sería el objetivo de la regresión logística. QUe siga nuestra ecuación.

Básicamente el objetivo final es traducirlo en una compra o compra.

En lugar de tener los datos originales como valores y como valores de predicción, se usará para predecir las probabilidades, y dicha probabilidad tiene que vivir entre entre cero y uno.

A la hora de hacer la predicción, lo que se busca es qué tan probable es que el suceso ocurra o no.

CLASIFICACIÓN

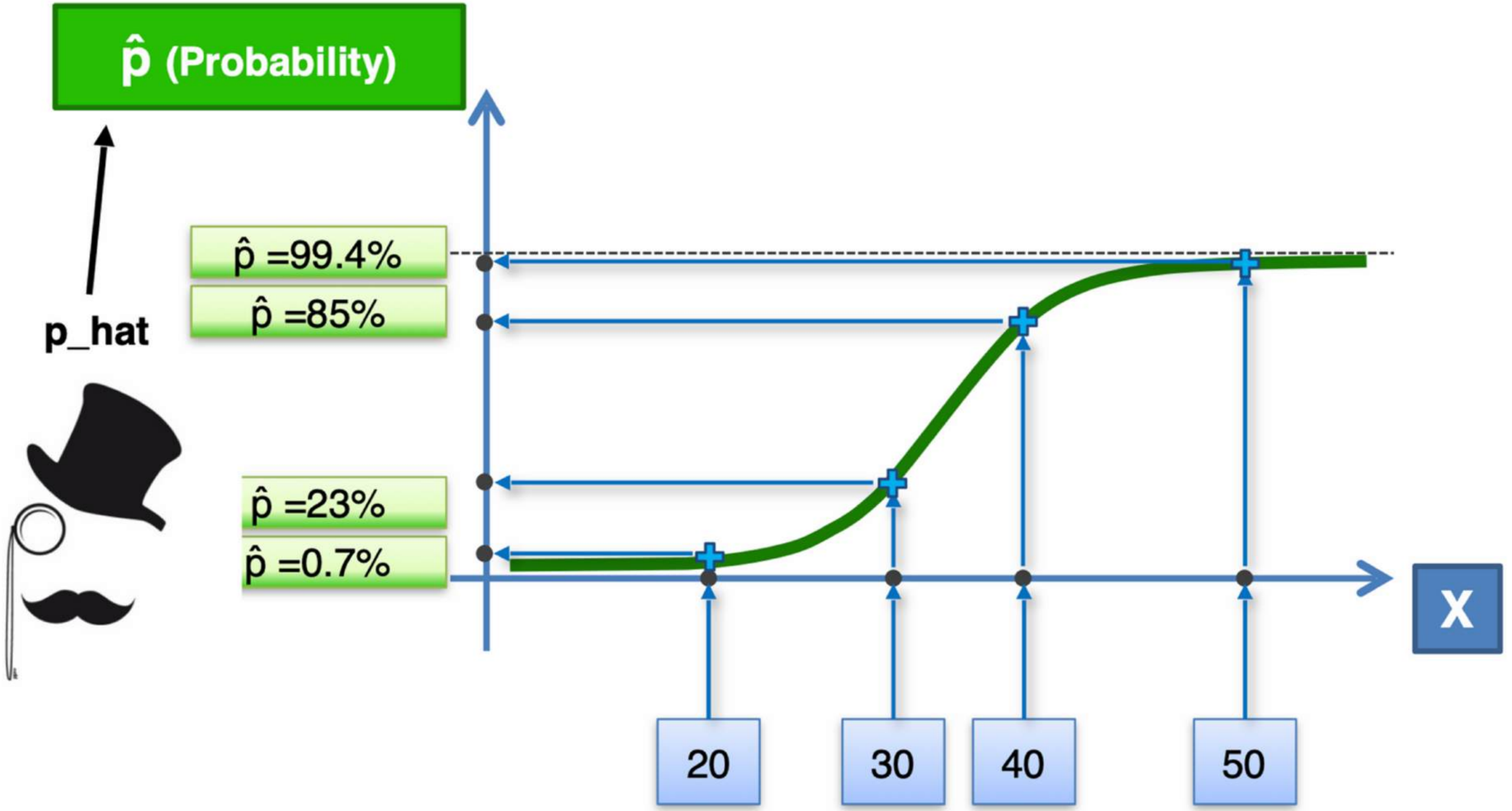
Para realizar el algoritmo hay que tomar una serie de valores aleatorios para la variable independiente x.
Vamos a predecir con edad 20, 30, 40 o 50 años cuál es la probabilidad de compra. Y estos valores se colocan en el eje X

Y ahora lo que hacemos es contrastar, encontrar la probabilidad de compra. Para ello, lo que hacemos es proyectar esos cuatro puntos en las observaciones de la curva logística.

Ahora, si proyectamos estos valores hacia el eje vertical, obtendríamos la probabilidad de compra.

El primer paso es obtener esa curva logística, ese cálculo de probabilidades. Una vez que tenemos esa curva que nos predice qué tan probable es que ocurra algo, usamos esa probabilidad como un score, como una puntuación que se puede utilizar a la hora de decidir la clasificación.

CLASIFICACIÓN



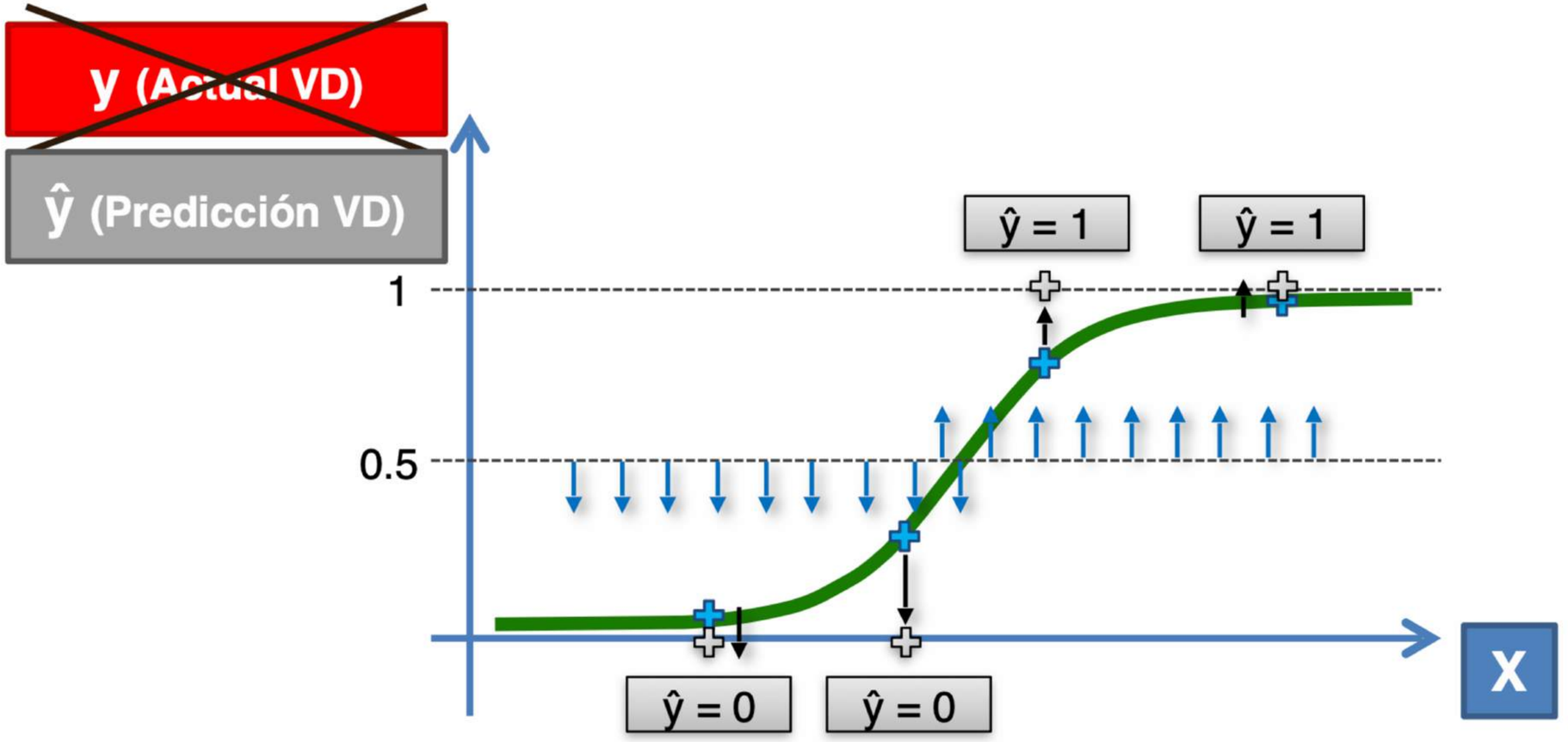
CLASIFICACIÓN

Evidentemente ese valor real no lo sabemos hasta que la persona no lleva a cabo la compra o la no compra. Pero nuestro modelo lo que hace es en lugar de hacer la predicción del valor real, hace una predicción del valor aproximado, es ese gorro.

De modo que el gorro que tenemos es una predicción de si la variable dependiente va a tomar el valor cero o valor uno. Es un enfoque un poquito arbitrario, pero básicamente esa es la idea.

Y en este caso se suele decidir que como valor central, el 0,5 marca la diferencia entre que sea más probable que compre o que no compre, que ocurra o no ocurra el suceso. Realmente esta línea es arbitraria. Cada quien puede colocar más arriba o más abajo.

CLASIFICACIÓN



DESCRIPCIÓN DEL PROBLEMA

REGRESIÓN LOGÍSTICA

El dataset a trabajar contiene datos acerca de redes sociales de publicidad en redes sociales. Cada una de las observaciones consta de un usuario en cierta red social.

Imaginen cualquier red social donde tenemos información del identificador de usuario del género de la persona. Por tanto, una variable categórica hombre o mujer, la edad del individuo, el sueldo estimado de la persona y si compraron o no compraron el producto que se ofertaba.

Pueden imaginar que se trata de Facebook y que es básicamente un anuncio en el lateral de la página de los que a veces hacemos clic para comprar un producto o para ver un curso online para cualquier tipo de cosa que pasa por publicidad.

REGRESIÓN LOGÍSTICA

Básicamente, esa sería la información que tiene este dataset junto con información sociodemográficas del género o la edad del individuo.

Para el propósito de nuestro análisis, imaginemos que una cierta compañía de coches acaba de lanzar su último coche de lujo, y básicamente se quiere anunciar en nuestra red social y acabar decidiendo si compra o no el usuario, el coche.

Entonces esa columna marcaría un cero si el determinado usuario decide no comprar ese coche debido al precio desorbitado.

Y en este caso tendríamos el valor uno si la persona al final decide comprar el coche.

REGRESIÓN LOGÍSTICA

La matriz de características va a constar únicamente de la edad y el sueldo estimado. Y la variable que vamos a predecir será el valor de compra o no compra.

Tenemos un total de 400 observaciones, por tanto, podemos recuperar esa división en conjunto de entrenamiento y conjunto de testing. Como 400 es un número redondo, podríamos elegir 300 para entrenar, 100 para testing.

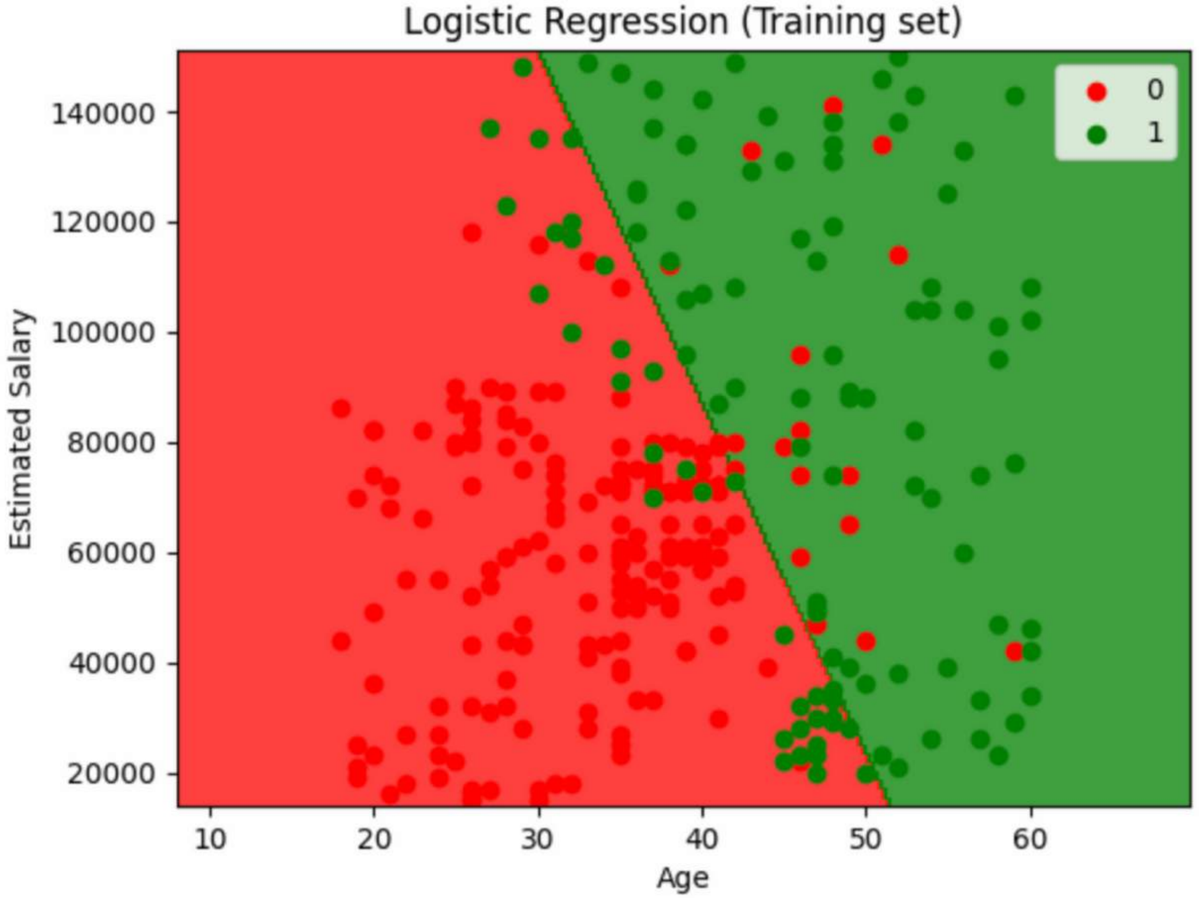
ALGORITMO

RETO

Reto:
Comencemos con:

- Active su ambiente local creado con miniconda en su equipo o su código de COLAB
- Abra el archivo 07_Simple_Logistic_Regression.ipynb
- Ejecute el código y analícelo junto al instructor

Resultado Esperado:



EVALUACIÓN DEL RENDIMIENTO

REGRESIÓN LOGÍSTICA

Queremos ver si esta predicción es o no es buena y evidentemente comparar elemento a elemento si la predicción ha sido la que esperábamos, no es la técnica recomendada.

Así que lo vamos a hacer es crear una matriz de confusión. Es una técnica muy potente. Se calcula la matriz de confusión sobre el conjunto de testing y podremos ver si las predicciones que ha elaborado nuestro algoritmo son potentes, nos sirven de algo para contrastar que el resultado realmente de la predicción casa con los datos que teníamos en el conjunto de test. Para ello se evaluarán cuantas son las predicciones correctas tanto de la categoría cero como de la categoría uno (compra o no compra) y en cuántas se ha equivocado el algoritmo.

Para crear esta matriz de confusión, realmente no tiene mucho misterio

EVALUACIÓN

PREGUNTAS Y RESPUESTAS

Mtro. Alfonso Gregorio Rivero Duarte

Senior Data Manager – CBRE

(+52) 5528997069

devil861109@gmail.com

https://www.linkedin.com/in/alfonso-gregorio-rivero-duarte-139a9225/