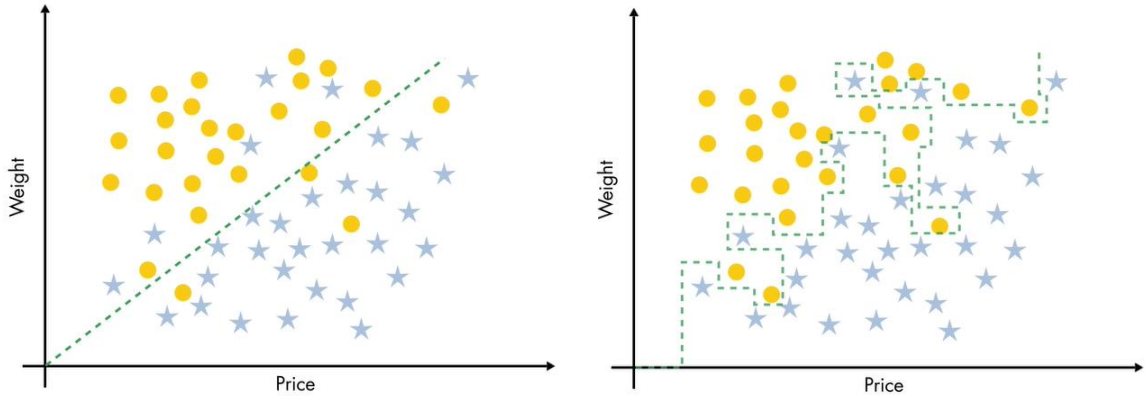
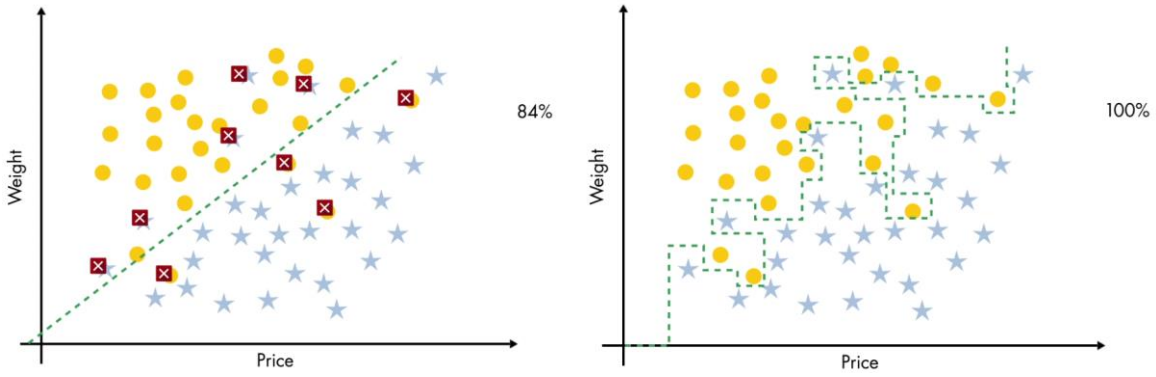
DATOS DE ENTRENAMIENTO Y PRUEBA

Aquí tenemos algunos datos con dos variables de predicción y dos clases de salida. ¿Cómo crearía un modelo de clasificación para estos datos? Instintivamente podría crear un modelo muy simple. Si el precio es superior al peso, una estrella. De lo contrario, un círculo. Pero hay otro modelo posible. Es un poco más complejo. ¿Qué modelo es mejor?

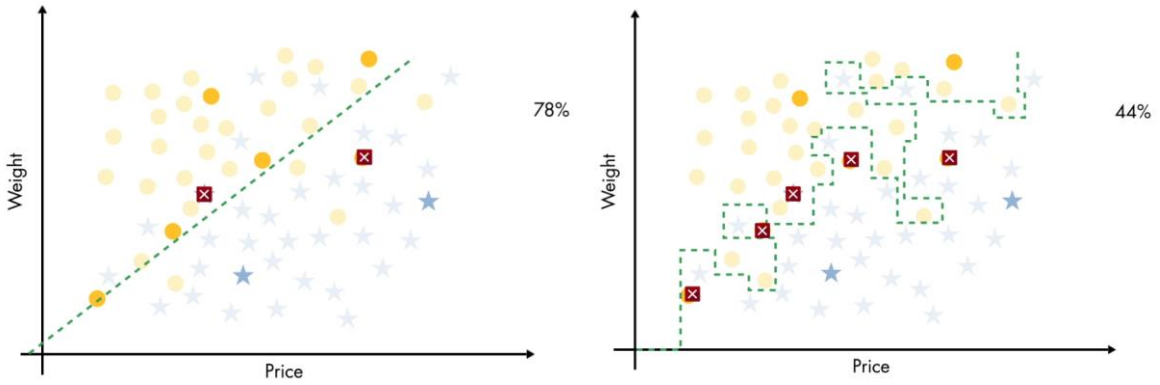


Bien, si los puntuamos según su precisión sobre los datos de entrenamiento, ganaría el segundo modelo. Es perfecto, mientras que el primer modelo genera algunas clasificaciones erróneas.



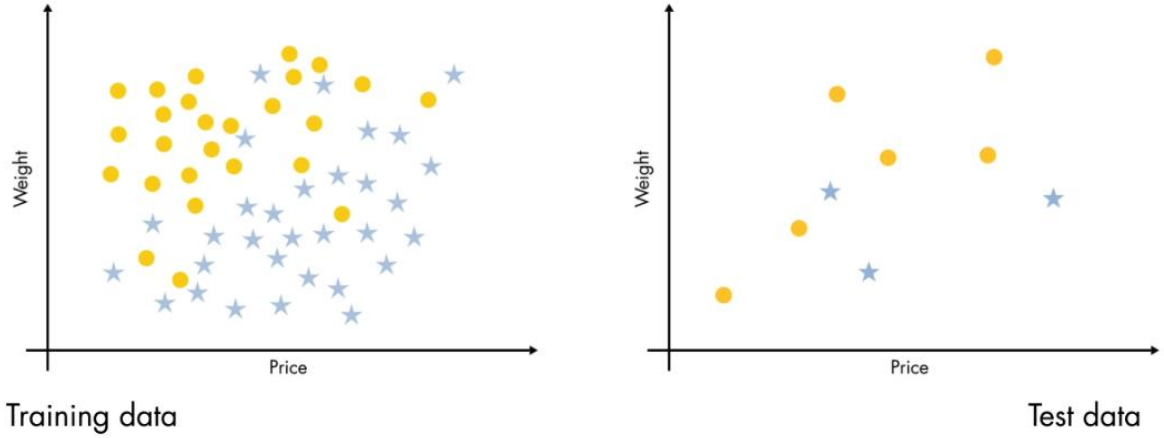
Pero, intuitivamente, parece que el segundo modelo es realmente demasiado complejo. Intenta capturar todos los detalles de los datos, incluso el ruido, mientras que el primer modelo se centra solo en las tendencias generales.

Si intenta usar estos modelos en los datos nuevos, probablemente verá que el modelo simple lo hace bien, mientras que el modelo complejo no es tan bueno. O, al menos, no tan bueno como prometía con esa puntuación perfecta.



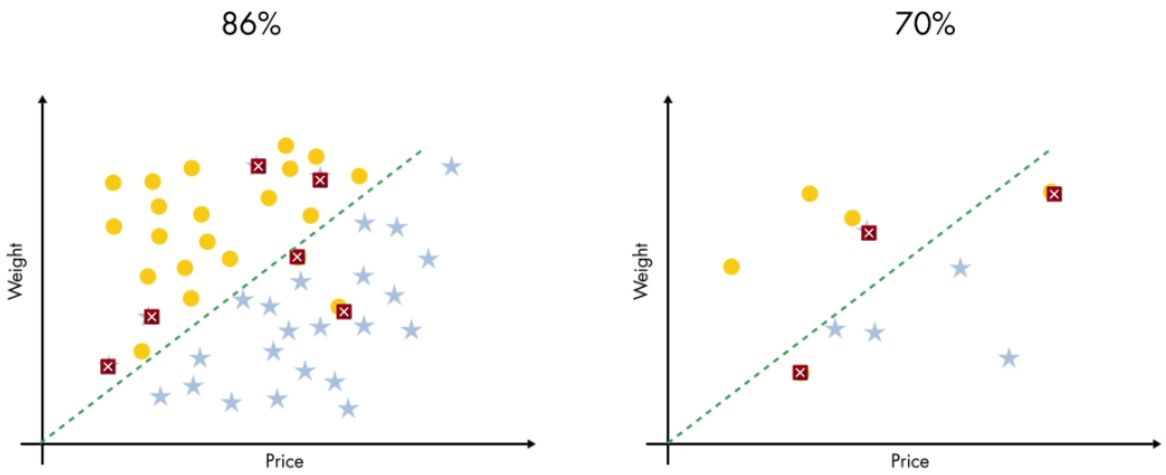
Este es un problema habitual en machine learning, conocido como sobreajuste: cuando el modelo lo hace muy bien con los datos usados para el entrenamiento, pero no generaliza bien con las nuevas observaciones.

El sobreajuste normalmente resulta obvio si puede probar el modelo proporcionándole algunos datos nuevos cuando sabe el resultado correcto. Pero ¿dónde se obtienen esos datos de prueba? En lugar de salir y obtener más datos, se suelen usar los datos que ya se tienen. Antes de entrenar el modelo, divida los datos en un conjunto de entrenamiento y otro de prueba.

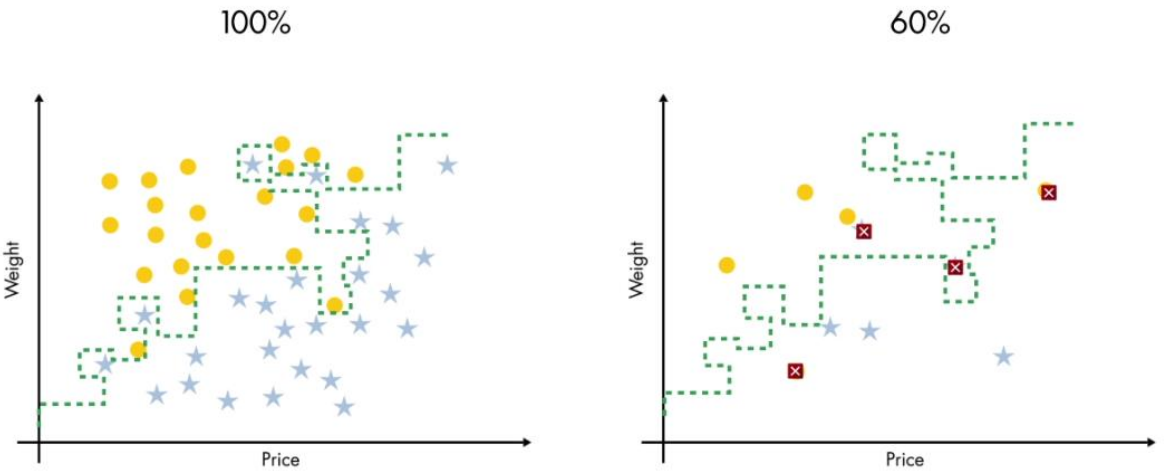


Hay distintas formas de hacerlo. Pero, la más simple, es seleccionar al azar una parte de las observaciones para reservarlas como conjunto de prueba. Después, el resto se usa para entrenar el modelo.

Vamos a probarlo con nuestro ejemplo. El modelo simple no cambia, aunque lo entrenemos con un conjunto diferente de observaciones. Sigue cometiendo algunos fallos. Y, cuando lo probamos con las nuevas observaciones, tiene un rendimiento similar. Esto es una buena señal.



Sin embargo, el modelo complejo depende mucho de los datos específicos que ve. Y, cuando lo probamos, vemos que lo hace peor. Eso, es el sobreajuste.



Esto confirma nuestra intuición de que el modelo simple es mejor, porque se ajusta al patrón general de los datos, no a los detalles. Siempre debe usar alguna forma de validación para ver cómo generalizan los modelos con datos nuevos, ya que, al final, eso es realmente lo que quiere que haga el modelo.