La evaluación del modelo nos dice cómo funciona nuestro modelo en el mundo real.

La evaluación dentro de la muestra nos dice qué tan bien nuestro modelo se ajusta a los datos ya proporcionados para entrenar. No nos da una estimación de qué tan bien el modelo entrenado puede predecir nuevos datos.

La solución es dividir nuestros datos, usar los datos In-sample o los datos de entrenamiento para entrenar el modelo.

El resto de los datos, denominados datos de prueba, se utilizan como datos fuera de la muestra.

Estos datos se utilizan luego para aproximar el rendimiento del modelo en el mundo real.

Separar los datos en conjuntos de entrenamiento y prueba es una parte importante de la evaluación del modelo.

Usamos los datos de prueba para tener una idea de cómo funcionará nuestro modelo en el mundo real.

Cuando dividimos un conjunto de datos, generalmente la mayor parte de los datos se usa para entrenamiento y una

la parte más pequeña se utiliza para la prueba.

Por ejemplo, podemos usar el 70% de los datos para entrenamiento; luego usamos el 30% para las pruebas.

Usamos un conjunto de entrenamiento para construir un modelo y descubrir relaciones predictivas.

Luego usamos un conjunto de prueba para evaluar el rendimiento del modelo.

Cuando hayamos completado la prueba de nuestro modelo, debemos usar todos los datos para entrenar el modelo.

Una función popular en el paquete de aprendizaje sci-kit para dividir conjuntos de datos es la "train\_test\_split".

Esta función divide aleatoriamente un conjunto de datos en subconjuntos de entrenamiento y prueba

La salida es una matriz: "x\_train" y "y\_train", los subconjuntos para el entrenamiento; "x\_prueba" y "y\_prueba",

los subconjuntos para la prueba. En este caso, el porcentaje de "tamaño de la prueba" de

los datos para el conjunto de prueba. Aquí está el 30%. El estado aleatorio es una semilla aleatoria para aleatorio

división de conjuntos de datos.

El error de generalización es una medida de qué tan bien nuestros datos predicen cosas nunca antes vistas.

datos.

El error que obtenemos usando nuestros datos de prueba es una aproximación de este error.

Esta figura muestra la distribución de los valores reales en rojo en comparación con los predichos

valores de una regresión lineal en azul.

Vemos que las distribuciones son algo similares.

Si generamos la misma gráfica usando los datos de prueba, vemos que las distribuciones son relativamente

diferente.

La diferencia se debe a un error de generalización y representa lo que vemos en la realidad.

mundo.

El uso de una gran cantidad de datos para el entrenamiento nos brinda un medio preciso para determinar cómo funciona nuestro modelo.

funcionará en el mundo real, pero la precisión de la actuación será baja.

Aclaremos esto con un ejemplo.

El centro de esta diana representa el error de generalización correcto; digamos nosotros

tome una muestra aleatoria de los datos usando el 90% de los datos para entrenamiento y el 10% para prueba.

La primera vez que experimentamos obtenemos una buena estimación de los datos de entrenamiento.

Si volvemos a experimentar, entrenando el modelo con una combinación diferente de muestras,

También obtendremos un buen resultado, pero los resultados serán diferentes en relación con la primera vez que

ejecutar el experimento.

Repitiendo el experimento nuevamente con una combinación diferente de muestras de entrenamiento y prueba,

los resultados son relativamente cercanos al error de Generalización, pero distintos entre sí.

Repitiendo el proceso, obtenemos una buena aproximación de la generalización

error, pero la precisión es pobre, es decir, todos los resultados son extremadamente diferentes de uno

otro.

Si usamos menos puntos de datos para entrenar el modelo y más para probar el modelo, la precisión de

el rendimiento de generalización será menor, pero el modelo tendrá buena precisión.

La figura de arriba demuestra esto; todas nuestras estimaciones de error están relativamente juntas,

pero están más lejos del verdadero rendimiento de generalización.

Para superar este problema, utilizamos la validación cruzada.

Una de las métricas de evaluación fuera de muestra más comunes es la validación cruzada.