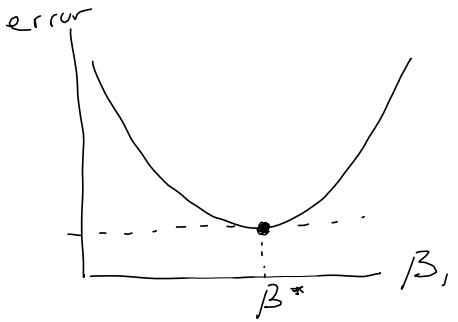
Métricas de evaluación regresión

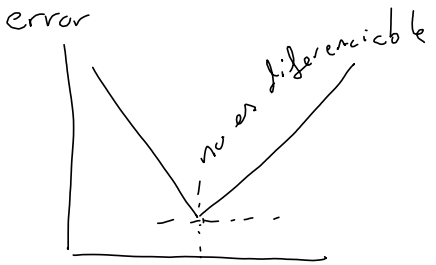
Nombre	Fórmulas $e = y - \tilde{y}$ i: filas	Ventajas	Desventajas
Mean squared error (MSE) $[0, +\infty)$ ✓ ✗	$\frac{1}{n}\sum_{i=1}^{n} e_i^2$	- misma que Reg. Lineal	- e^2 perdemos magnitud error
Root MSE $[0, +\infty)$ ✓ ✗	$\sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$	- magnitud es más comparable con el target	- no es el error medio exacto
Mean absolute error (MAE) $[0, +\infty)$ ✓ ✗	$\frac{1}{n}\sum_{i=1}^{n} \lvert e_i \rvert$	- visión exacta del error medio	- Reg lineal no se ajusta con el MAE.
Mean absolute % error $[0, +\infty)$ ✓ ✗	$\frac{100\%}{n}\sum_{i=1}^{n} \left\lvert \frac{e_i}{y_i} \right\rvert$	- error relativo	- no funciona si target cercano 0
R^2 $[0, 1]$ ✗ ✓	$1 - \frac{\sum_i (y_i - \tilde{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$	- comparable con otros datasets	- no da magnitud del error

---

note: $\min_{\beta_1} \sum_{i=1}^{n} e_i^2$



$\min_{\beta_1} \sum_{i=1}^{n} |e_i|$



Workflow

Data Understanding → Data Preparation → Modelling → Evaluation

Data Understanding

→ relación atributo y target → Modelling / Evaluation

→ entender atributo y posible limpieza → Data Preparation

Data Preparation

→ "traducir" todas las variables a numéricas

- → Fechas: extraer inputs mes, año, día, trimestre, ...
- → nulos: 1. Tengo info en otra variable, 2. Imputación inteligente, 3. Imputación "a saco" (media, moda, ...)
- → categóricas
  - → Boolean → orden intrínseco (versiones) → orden de fuentes externas → OHE (<100) → agrupar → Frec. encoding

→ preparar según el algoritmo (e.g. outliers)

- → Reglas: all in, sin cambios
- → Geometricos: transformación logarítmica de target y atributo numérico (relación lineal)
- → Vecindad: scaling

Modelling

1. Instanciar: escoger algoritmo y parámetros
2. Entrenar: .fit
   └ X_train, y_train

Evaluation

1. Overfitting: comparar (y_train, pred_y train) vs (y_test, pred_y_test)

   Metricas dependen de la tarea

   Clasificación: Accuracy, F1 Score, Confusion matriz, AUC

   Regresión: MSE, RMSE, MAE, MAPE, R^2

2. Desbalanceo: objetivo NO es tener proporcion igualitaria

   → Oversampling → prediga mejor clase minoritaria

   → Undersampling

   development, NO validation

¿Porqué mi modelo funciona mal?

1. Memorización: el modelo contiene reglas que solo aplican a los datos observados
2. Realidad tiene una distribución distinta

→ realidad: decisión consciente

train / test aleatoria: quiero misma distribución train / test

- random hold out: una sola partición
  - si tengo pocos datos: múltiples versiones
    - K-Fold Cross Val.
    - Bootstrap