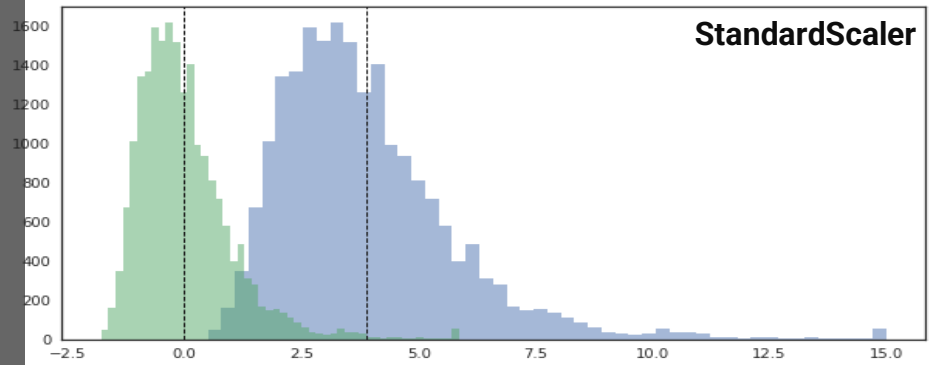
Методи підготовки, створення та відбору признаків для тренування моделі

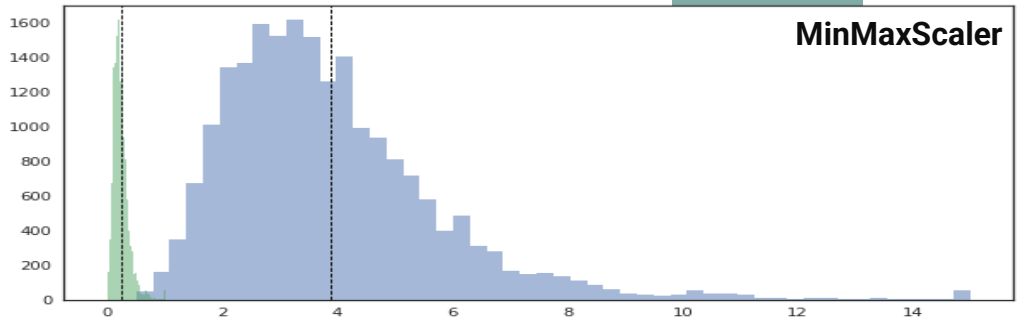
1. Стандартизація та масштабування
2. Кодування признаків (One-Hot, Label, Target)
3. Проблеми, пов’язані із незбалансованістю спостережень та методи їх усунення
4. Вибір признаків
5. Підготовка датасетів для навчання моделей (Валідаційні датасети, крос-валідація)
6. Метрики оцінки якості моделей
7. Перенавчання та недонавчання моделей. Способи боротьби із недо- та перенавчанням

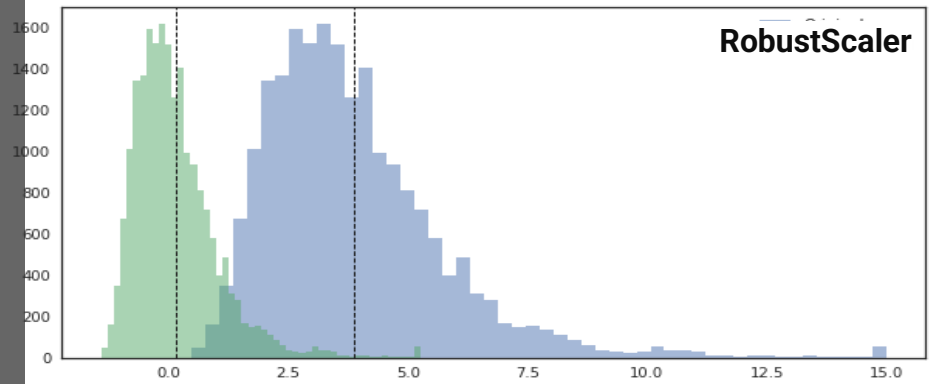
Стандартизація та масштабування

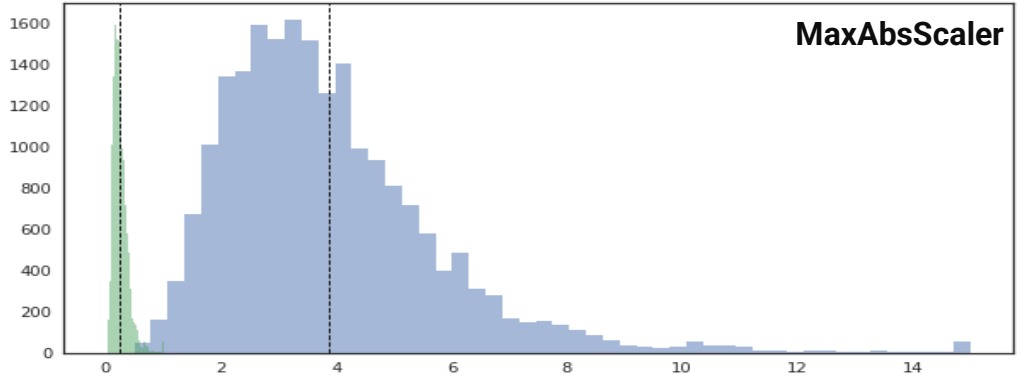
Назва методу	Формула	Результат	Умови використання
StandardScaler	$X^{std} = \frac{(X - \bar{X})}{\sigma_x}$	Середнє значення 0, стандартне відхилення 1	Нормальний закон розподілу
MinMaxScaler	$X_{\text{minmax}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$	Масштабовані значення знаходяться в межах бажаного діапазону, зазвичай [0, 1]	Необхідність забезпечити певні межі діапазону змінних та зберегти форму розподілу (використовується майже завжди)
RobustScaler	$X_{\text{robust}} = \frac{X - \text{median}(X)}{Q_3 - Q_1}$	Масштабування засновано на значеннях міжквартільного діапазону (IQR). Викиди мають менший вплив	Використовується коли є певні викиди в даних
MaxAbsScaler	$X_{\text{maxabs}} = \frac{X}{X_{\text{absmax}}}$	Масштабовані значення знаходяться в межах діапазону [-1, 1]	Використовується коли дані мають середнє значення близько 0, розріджені дані

Порівняння ефекту різних методів масштабування









Кодування признаків (One-Hot, Label, Target)

Метод	Умови використання	Результат
One-Hot	1. Категоріальні дані, які не мають природної впорядкованості 2. Кількість категорій відносно мала	Вектор бінарних змінних
Label	1. Категоріальні дані, які мають природну впорядкованість 2. В подальшому використовується методи на основі дерев	Цілі числа замість категорій
Target	Велика кількість унікальних значень категорій	Вектор ймовірностей (частот) спостереження кожного класу для кожного унікального значення категорії, або середнє значення залежної змінної

Порівняння методів кодування

One-Hot

	Area	Year	Item	Cassava	Maize	Plantains and others	Potatoes	Rice, paddy	Sorghum	Soybeans	Sweet potatoes	Wheat	Yams
0	Albania	1990	Maize	0	1	0	0	0	0	0	0	0	0
1	Albania	1990	Potatoes	0	0	0	1	0	0	0	0	0	0
2	Albania	1990	Rice, paddy	0	0	0	0	1	0	0	0	0	0
3	Albania	1990	Sorghum	0	0	0	0	0	1	0	0	0	0
4	Albania	1990	Soybeans	0	0	0	0	0	0	1	0	0	0

Target encoding

	Area	Year	Item	Items_target_enc
0	Albania	1990	Maize	36310.07
1	Albania	1990	Potatoes	199801.55
2	Albania	1990	Rice, paddy	40730.43
3	Albania	1990	Sorghum	18635.78
4	Albania	1990	Soybeans	16731.09

Label Encoding

	Area	Year	Item	Items_LE
0	Albania	1990	Maize	3
1	Albania	1990	Potatoes	9
2	Albania	1990	Rice, paddy	4
3	Albania	1990	Sorghum	1
4	Albania	1990	Soybeans	0

Проблеми незбалансованості спостережень (для задач класифікації)

Незбалансованість спостережень - наявність у зібраних даних одного або декількох класів, кількість спостережень яких суттєво менша за інші

Суть проблеми	**Наслідки проблеми**
Неточність моделі	Низька якість моделі при виявлені “рідких” класів. Неспроможність розпізнавати рідкий клас
Низька “узагальнююча здатність” моделі	Низька якість моделі на нових даних
Некоректність показників якості моделі	Найкраща точність та похибка моделі досягається наївним прогнозом найбільшого класу
Труднощі в інтерпретації моделі	Відсутність достатньої кількості спостережень не дозволяє коректно інтерпретувати причину помилкової класифікації
Некоректна інтерпретація “важливості” незалежних змінних	Змінні, що мають велику предиктивну здатність для рідких класів можуть помилково вважатися неважливими
Чутливість до розбиття датасетів	Для різних способів розбиття якість моделі суттєво змінюється

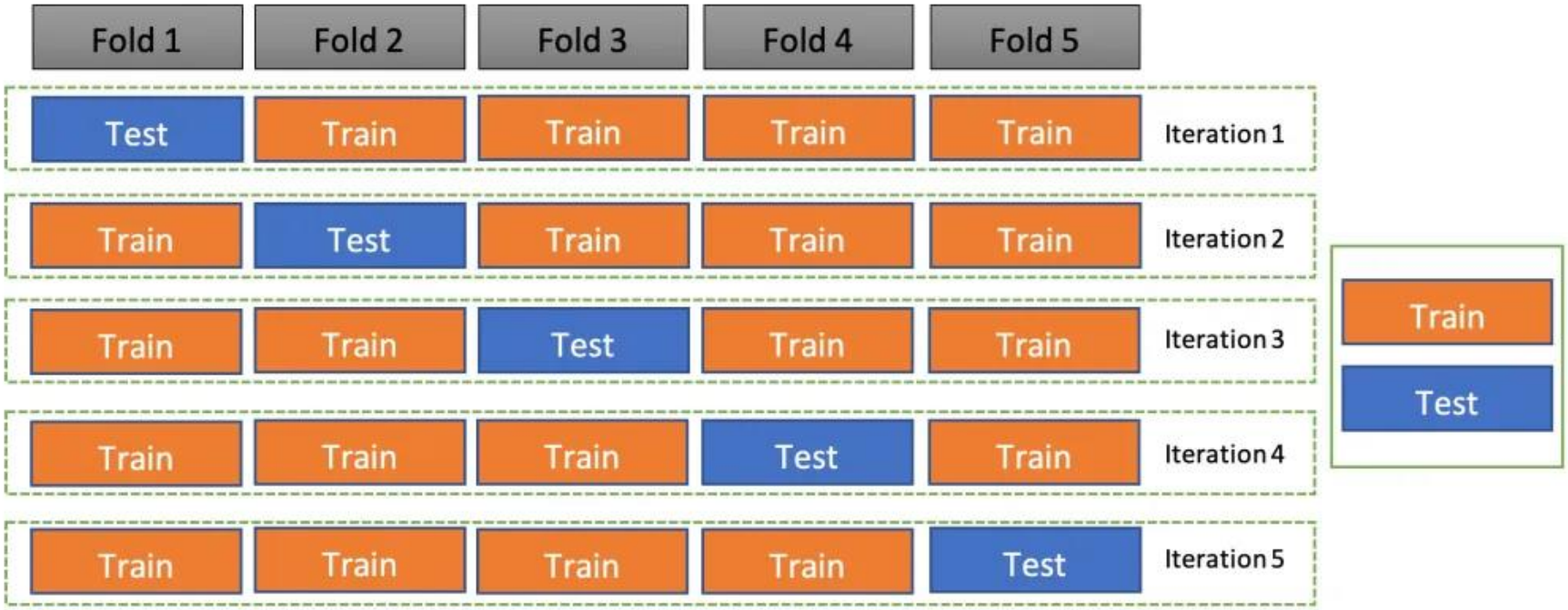
Способи підготовки та навчання моделей на незбалансованих даних

Метод	Способи втілення
Перебалансування вхідних даних	1. Зменшення кількості спостережень масового класу (випадкова вибірка), 2. Синтез “нових” спостережень малого класу (SMOTE): 2.1. Для кожного спостереження із малого класу вибрати k найближчих сусідів 2.2. Для кожного спостереження вибрати “випадкового” найближчого сусіда 2.3. Сформувати “синтетичне” спостереження як суперпозицію оригінального спостереження та вибраного найближчого сусіда 2.4. Повторити кроки 2.2. та 2.3.
Зважування помилки в процесі навчання	Помилка класифікації для малого класу встановлюється більшою ніж для масового класу
Вибір спеціальних метрик оцінки якості моделі	Використання таких метрик як Precision, Recall, F1-score, Area Under the ROC Curve (AUC-ROC)

Методи вибору признаків для побудови моделі

Метод	**Способи втілення**
Фільтрація	прибирання ознак із майже унікальними або майже постійними значеннями
Статистичний відбір	Для задач регресії: • Коефіцієнт кореляції Пірсона (`f_regression`) • Mutual information (`mutual info regression`) $\mathrm{MI}(X,Y) = \iint p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dx\, dy$ Для задач класифікації: • Chi-2 тест (`chi2`) • ANOVA F-значення (`f_classif`) • Mutual information (`mutual info classif`) $I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$
Вибір на основі моделі	L1 регуляризація, Feature Importance, Feature Permutation

Підготовка датасетів для навчання моделей



Метрики оцінки якості моделей (регрессія)

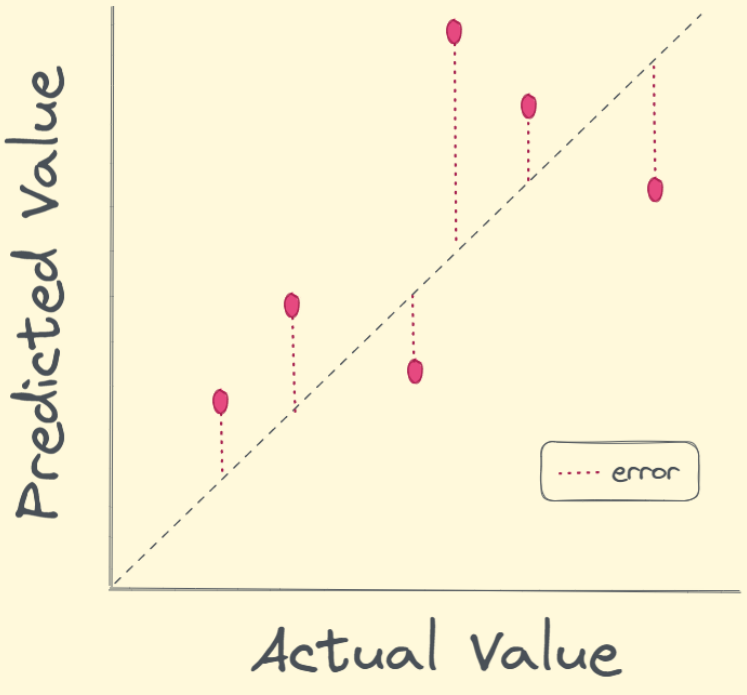
$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

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

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

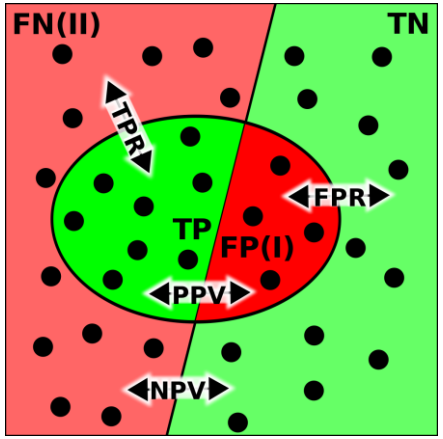
$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$



Метрики оцінки якості моделей (класифікація)

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP



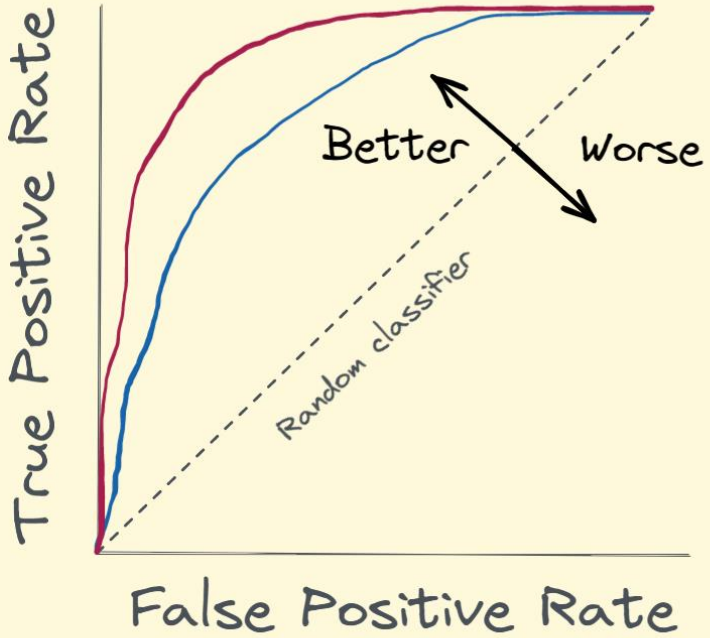
$$\text{Accuracy} = \frac{TP + TN}{\text{Total Samples}}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

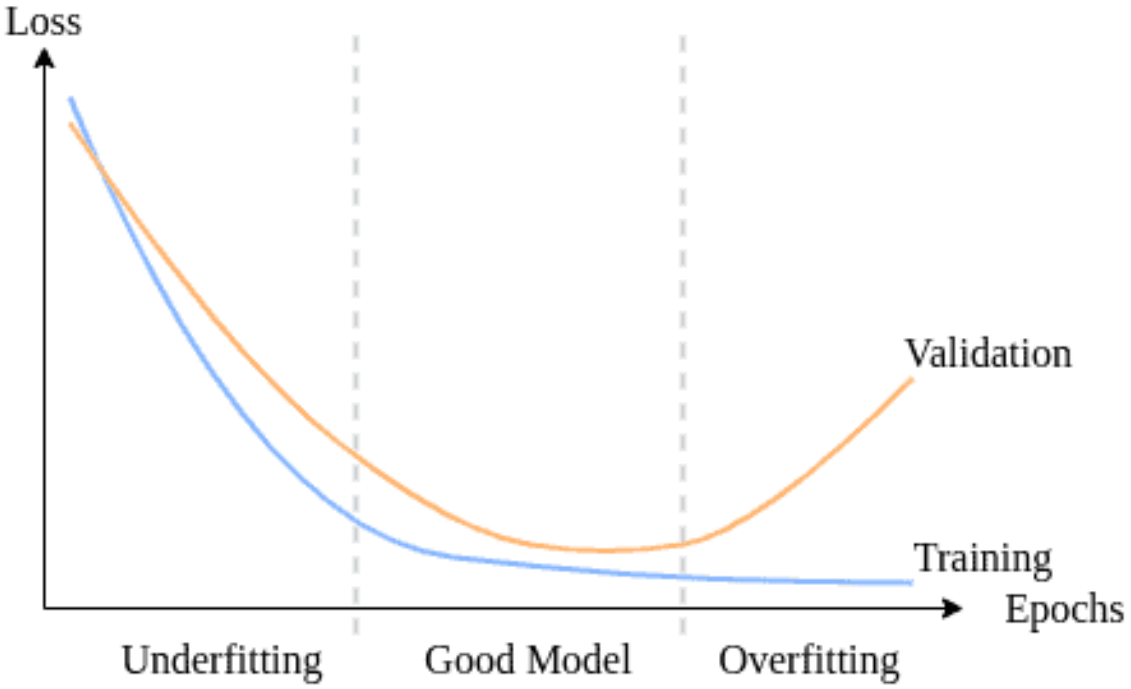
$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Log Loss} = -\sum_{k}^{K} y^{(k)} \log(p^{(k)})$$



Перенавчання та недонавчання моделей



Перенавчання та недонавчання моделей

Недонавчена модель	Перенавчена модель
Модель надпроста	Модель надскладна
Великі похибки на train даних	Дуже точна на train даних
Великі похибки на тестових даних	Великі похибки на тестових даних
Треба додати параметрів, факторів в модель	Треба прибрати параметри, фактори з моделі
Зменшити вплив регуляризації	Збільшити вплив регуляризації
Збільшити тривалість навчання	Зменшити тривалість навчання
Додати даних для навчання моделі	

Питання?