

Pattern Recognition and Machine Learning

Lecture 06 Feature Engineering

Chizhi Chris ZHANG

zhangchizhi@ciomp.ac.cn

Advanced Computing and Digital Technology Research Center

University of Chinese Academy of Sciences



Autumn 2026

Lecture plan and time allocation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Topic	Minutes
Feature transformation	45
Feature construction	22.5
Filter selection	22.5
Wrapper and embedded selection	45

Three training rows and one new row

Save training transformations; construct useful quantities; select features honestly.

The numerical steps are shown on the slides.



One small table will anchor the calculations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A transformation has a training step and an application step. Estimate its rules on training rows, save them, and apply them consistently to new rows.

The learner receives a matrix, so each input column needs a defined meaning. Missing measurements, units and category codes change what the model can learn.

$$X \in \mathbb{R}^{n \times p}$$

Use measurements available at admission to predict a later outcome. Training has 3 rows. A new row is age 50, lab missing, ward C.

Rows	Age(years)	Lab(mg/L)	Ward	y
A	20	1	A	0
B	40	3	B	1
C	60	missing	A	1

Check. Discharge status is measured after the prediction time and cannot be an admission feature. A column can contain valid values yet be invalid for the task.

Fit learns a rule, transform applies that rule

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

Each fold fits preprocessing on its training rows and transforms its validation rows. A pipeline preserves this order.

$$50 - 40 = 10$$

Check. Fitting a scaler or supervised selector on all rows before validation lets validation information influence training.

Operation	Rows used	Saved or applied result
Fit centering	Training: 20, 40, 60	Save mean 40.
Transform a new row	New value 50	Use $50 - 40 = 10$.
Refit on a new batch	Would create another coordinate system	Do not silently replace the training rule.
Cross-validation	Repeat fitting inside each training fold	Apply that fold extquotesingle s rule to its validation rows.

Mean and median encode different assumptions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Count missingness by column and relevant groups. Distinguish unavailable, unmeasured and invalid values.

Observed values 1, 2, 3, 100 give mean 26.5 and median 2.5. Replacing a missing value by the mean places it far from the typical small values.

Check. Simple imputation reduces variation and ignores uncertainty. It is a baseline, not a recovery of the true missing value.



The training median fills the new row

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

(1, 3, missing). Fit the median using observed training values, $m = (1 + 3)/2 = 2$.

Worked calculation

$$m = (1 + 3)/2 = 2$$

$$\text{impute}(1, 3, \text{missing}) = (1, 3, 2), \quad \text{impute}(\text{new missing}) = 2.$$

$$q = \mathbf{1}[\text{lab missing}]$$

Check. An ordering policy change can alter what missingness means. The indicator is useful only if the process remains relevant.

Distribution and units guide the next operation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Check units, plausible ranges, skew, extreme values and subgroup mixtures. A large value can be a rare valid case or an entry error.

Check. A Gaussian-looking histogram is not itself evidence of useful predictive information. Keep rare valid cases unless the task provides a reason to remove them.

Observation	Possible explanation	Check before transforming
Height 170 among values near 1.7	Centimeters mixed with meters	Recover a common unit.
Long right tail	Large valid measurements or entry errors	Inspect records; compare a justified transform.
Two peaks	Different subgroups	Check group membership before forcing one shape.
An extreme but valid case	The event the task needs to detect	Preserve its meaning in the evaluation.

Standardization makes the scale explicit

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The new row supplies the numerator's raw measurement. It does not refit either saved quantity.

$$\mu = (20 + 40 + 60)/3 = 40, \quad s^2 = \frac{(-20)^2 + 0^2 + 20^2}{3} = \frac{800}{3}, \quad s \approx 16.33.$$

$$z_{\text{new}} = \frac{50 - \underbrace{40}_{\text{saved mean}}}{\underbrace{16.33}_{\text{saved scale}}} \approx .612.$$



A value 0.612 is 0.612 training standard deviations above the training mean. The transformation is affine and preserves ordering.

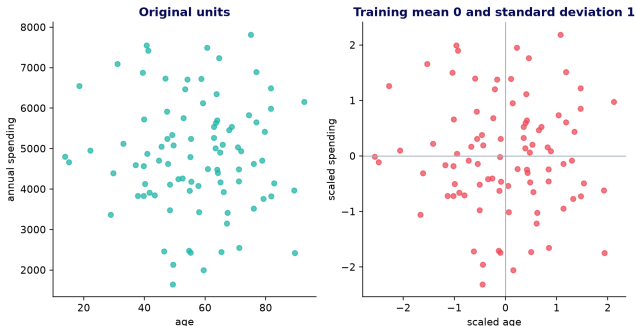
$$z_j = \frac{x_j - \mu_j}{s_j}, \quad \mu_j = \frac{1}{n} \sum_i x_{ij}, \quad s_j = \sqrt{\frac{1}{n} \sum_i (x_{ij} - \mu_j)^2}.$$

$$\mu = 40 \quad s = \sqrt{800/3} \approx 16.33$$

Scaling changes units while keeping point identities

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The same observations remain after scaling. Each column uses its own training center and scale. Standardization changes coordinates without creating new samples or guaranteeing a Gaussian distribution.



Scale changes distances and regularized scores

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

With age scale 10 years and lab scale 100 mg/L, the squared standardized distance is $1^2 + 1^2 = 2$.

$$10^2 + 100^2 = 10100 \quad 1^2 + 1^2 = 2$$

Check. A strictly increasing rescaling preserves numeric ordering and possible training partitions for an ideal single-feature threshold tree. This does not make scaling irrelevant for every pipeline.

Feature difference	Raw squared contribution	After dividing by its scale
Age: 10 years	$10^2 = 100$	$(10/10)^2 = 1$
Lab: 100 mg/L	$100^2 = 10000$	$(100/100)^2 = 1$
Total	10100	2

Minmax and robust scaling answer different needs

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Training range 20–60 maps 50 to 0.75. New 80 maps to 1.5 unless explicitly clipped. Refit-free extrapolation is expected.

Check. Robust scaling does not delete outliers. Minmax does not guarantee every future value remains within 0–1. Specify zero-range or zero-IQR handling.

Rule	Training information	What happens to a new value?
Minmax	Minimum 20, maximum 60	$50 \mapsto (50 - 20)/40 = .75$; $80 \mapsto 1.5$.
Robust	Training median and IQR	Center by the median, divide by the IQR.
Boundary handling	A zero range or zero IQR	Specify the handling; avoid refitting on the new batch.

Row normalization changes a different quantity

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Column scaling estimates a centre and scale for each feature across training rows. Row normalization divides one observation by its own length:

$$x \mapsto x / \|x\|_2.$$

$$\|(3, 4)\|_2 = 5$$

Check. Magnitude can matter for dose or spending. Handle zero vectors and avoid combining incompatible units without justification.



A log transform expresses diminishing changes

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The raw increases 9 and 90 become equal increments in transformed coordinates. Large values compress more strongly.

$$z = \log(1 + x), \quad x \geq 0 \text{ for the simple count example .}$$

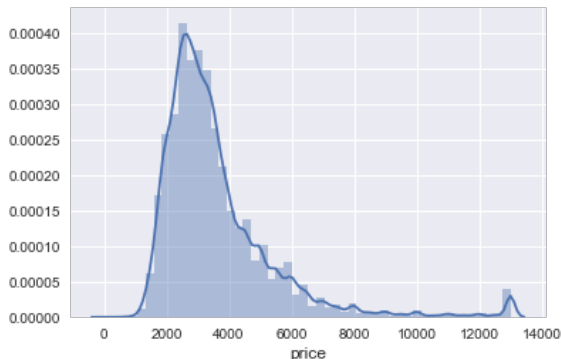
x	0	9	99
$\log(1 + x)$	0	$\log 10 \approx 2.303$	$\log 100 \approx 4.605$

Check. Negative measurements need a justified transform. Scaling and distribution symmetry do not by themselves establish a better predictive relationship.

Inspect a long right tail before transforming

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The reference shows the original price distribution and its long right tail. A log transform compresses large values. Inspect the transformed data before deciding whether this serves the task.



Binning turns a numeric scale into interval features

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Age bins $[0, 30)$, $[30, 50)$, $[50, \infty)$ map 20, 40, 60 to three different categories. New 50 enters the third bin.

$[0, 30)$, $[30, 50)$, $[50, \infty)$

Check. Binning loses within-bin differences and creates abrupt boundaries. More bins increase dimensions and can leave sparse intervals. Compare it with a continuous baseline.

Age	Saved interval	What the model retains
20 or 29	$[0, 30)$	The same bin identity; their exact difference is lost.
40	$[30, 50)$	The middle bin.
50 or 60	$[50, \infty)$	The third bin, because the left boundary is included.

One-hot encoding avoids an invented category order

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Codes 1, 2, 3 would imply numeric distances and an order. One-hot columns let a model learn separate category effects.

Ward	Indicator A	Indicator B	Indicator C
A	1	0	0
B	0	1	0
C	0	0	1

Check. For an unregularized linear model with an intercept, all category indicators are linearly dependent. Dropping one reference category can resolve this. Different regularized models can use all indicators.

Known categories and an unseen category

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Training categories are A and B, so the model expects exactly two indicator columns.

Input	A column	B column	Interpretation
A	1	0	Known category A
B	0	1	Known category B
New C	0	0	Unknown, if ignored explicitly

Creating a third column at prediction time would change the model input width. An explicit unknown category is another design choice.

Frequency encoding uses training counts, such as $8/10 = .8$ and $2/10 = .2$. Target encoding instead uses target outcomes and needs a stricter information boundary.



Target encoding must exclude a row's own answer

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A target statistic

For category k , let n_k be its training count and $\bar{y}_k$ its mean target. Shrink toward the overall training mean $\bar{y}$:

$$e_k = \frac{n_k \bar{y}_k + \alpha \bar{y}}{n_k + \alpha}.$$

For $n_k = 2$, $\bar{y}_k = 1$, $\bar{y} = .2$, $\alpha = 3$,
 $e_k = (2 + .6)/5 = .52$.

Whose labels may be used?

A unique ID with target 1 has a raw category mean of 1. Feeding that mean back to the same row reveals its answer.

Training rows use out-of-fold statistics. New rows use saved training statistics. Unseen categories can fall back to the training global mean.

Two different controls

Smoothing reduces unstable counts. Excluding the row's own target prevents leakage. Smoothing alone cannot replace exclusion.

Distance replaces four coordinates with a task quantity

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A learner can only use the coordinates it receives.

A domain hypothesis motivates the construction. Units, zero denominators and availability define whether it is valid.

Check. A sensible story is a reason to test a feature, not proof that it improves prediction.

On a local planar map in km, pickup is (1, 2) and destination is (4, 6).

$$d = \sqrt{(4 - 1)^2 + (6 - 2)^2} = \sqrt{9 + 16} = 5 \text{ km.}$$

Check. Straight-line distance differs from road distance. Latitude/longitude need spherical or projected geometry, not Euclidean distance in degrees.



Ratios and differences represent different hypotheses

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A ratio expresses size relative to a baseline. A difference retains an absolute unit. Keeping raw variables can retain magnitude lost by a ratio.

$$(6 - 4)/4 = 0.5$$

Check. A change from 0.01 to 0.03 is a 200% increase but only 0.02 in absolute units. Use justified denominator handling and inspect noise sensitivity.

Change	Absolute difference	Relative change
4 to 6	$6 - 4 = 2$	$(6 - 4)/4 = .5 = 50\%$
.01 to .03	$.03 - .01 = .02$	$.02/.01 = 2 = 200\%$
Interpretation	Retains an absolute unit	Depends strongly on the baseline denominator.

Feature expansion grows faster than the evidence

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Adding x_1x_2 lets the effect of one feature depend on the other. This expands the model representation and requires evidence for the added interaction.

$$\hat{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{12}x_1x_2.$$

$$\partial\hat{y}/\partial x_1 = \beta_1 + \beta_{12}x_2 \quad \mathbf{z} = \mathbf{x}_1 + \mathbf{x}_2 - 2\mathbf{x}_1\mathbf{x}_2$$

Check. An interaction coefficient is a model term. It does not by itself establish that changing a cause would change an outcome.

For two inputs, quadratic expansion gives the original features, their squares and one interaction. More inputs create many more columns, increasing the need for validation.

$$p + \frac{p(p+1)}{2} \quad \text{columns before the intercept.}$$

$$p = 2 \quad p = 100$$

$$100 + 5050 = 5150$$

Check. Many combinations may be poorly observed. Construct plausible features and compare regularized baselines.

Periodic time should preserve nearby hours

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A date can describe a trend; days since an event measure recency; hour of day describes a cycle.
Encode a 24-hour cycle as

$$\mathbf{z}_1 = \sin(2\pi h/24), \quad \mathbf{z}_2 = \cos(2\pi h/24).$$

Check. The same hour on different days maps to the same pair. Keep date/recency when needed, and compute elapsed time from the prediction cutoff.



A useful construction must survive an ablation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A seven-day purchase count uses events in $[t - 7 \text{ days}, t)$ that were recorded by time t .

Retrain with and without the constructed feature on identical folds. Compare the task metric, variation and acquisition cost.

Features used	Validation score	First interpretation
Original Columns	0.80	Compare with baseline
Original columns plus distance	0.84	May capture a useful relationship
Also add object ID	0.85	Check for memorized repeated objects

Check. These numbers illustrate comparison logic. An ID improvement may reflect memorized entities rather than transfer to new entities.

Selection keeps columns, extraction creates coordinates

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Construction can produce many correlated or weak columns.

Selection retains measurements. Extraction combines them into coordinates. Construction adds a variable based on a task hypothesis.

Operation	Output example	Consequence
Selection	Keep age and lab	Retain measurement meanings
Extraction	$z_1 = 0.7x_1 + 0.7x_2$	Combine measurements
Construction	Price/quantity	Test a task hypothesis

Check. PCA is an extraction method, treated in Lecture 11. Removing variables is not the same operation as projecting to principal components.

Variance and correlation catch simple redundancy

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Every fitted selector belongs inside the training part of validation. A supervised filter is still using labels even before the final model fits.

A column with the same value in every training row has variance 0 and cannot separate those rows. A near-constant column may still mark an important rare event.

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}.$$

$$y = x^2$$

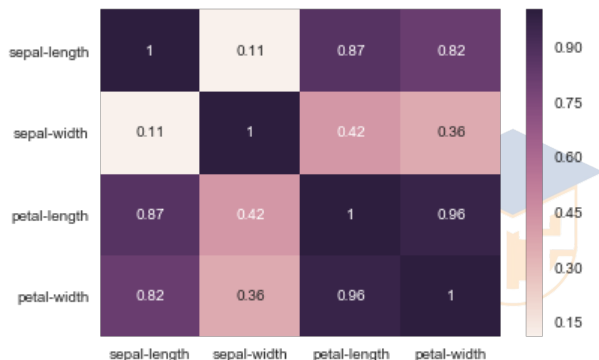
Check. Correlation measures linear association. $y = x^2$ on symmetric inputs can have correlation 0. Unsupervised redundancy removal still needs a justified choice of which measurement to retain.



Read one pair of features in a correlation matrix

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

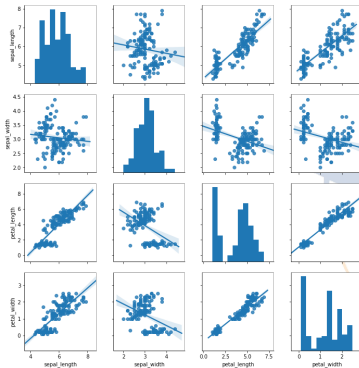
Petal length and width have correlation near .96. The diagonal entries compare each feature with itself. Strong correlation suggests redundancy, but it cannot alone determine which feature to remove.



A pairplot shows distributions and dependencies

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Sepal length, sepal width, petal length, petal width. Rows and columns use this same order.



A filter asks about one column at a time

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

In XOR, either input alone has zero label information, while their pair determines the label. A low marginal score can miss a useful combination.

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

 $x_1 = 0$ $x_1 = 1$

$$y = x_1 + x_2 - 2x_1x_2.$$



Mutual information measures label dependence

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Information gain

For discrete variables, mutual information is the reduction in label uncertainty after observing a feature.

$$I(X; Y) = \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} = H(Y) - H(Y | X).$$

Four equally likely rows	$H(Y)$	$H(Y X)$	$I(X; Y)$
$Y = X$: (0, 0), (0, 0), (1, 1), (1, 1)	1	0	1 bit
Independent: (0, 0), (0, 1), (1, 0), (1, 1)	1	1	0 bits



In the first case, knowing X fixes Y . In the second, either value of X still leaves a 50–50 label.

Dependence can be nonlinear: $Y = X^2$ may carry information even with zero linear correlation. Continuous data require an appropriate estimator fitted on training rows.

Chi-square compares observed and expected counts

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Among 100 cases, 40 have the feature and 40 are positive. Expected present/positive count is $40 \cdot 40/100 = 16$, while 30 are observed.

$$O = \begin{bmatrix} 30 & 10 \\ 10 & 50 \end{bmatrix}, \quad E = \begin{bmatrix} 40(40)/100 & 40(60)/100 \\ 60(40)/100 & 60(60)/100 \end{bmatrix} = \begin{bmatrix} 16 & 24 \\ 24 & 36 \end{bmatrix}.$$

$$\begin{aligned} \chi^2 &= \sum_{\text{four cells}} \frac{(O - E)^2}{E} \\ &= \frac{14^2}{16} + \frac{(-14)^2}{24} + \frac{(-14)^2}{24} + \frac{14^2}{36} \approx 34.028. \end{aligned}$$



Check. Add discrepancies from all four cells. A large score challenges independence, but does not establish causation. Sklearn's chi-square selector requires nonnegative inputs, so standardized negative values violate that interface.

Subset selection uses a predictor as its judge

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A candidate subset or penalty is itself a modeling choice.

Which subset predicts well with this chosen learner? For each candidate subset, fit the predictor on training rows and score separate validation rows.

$$p = 20$$

Check. Trying many subsets can exploit validation noise. Use an outer evaluation loop or an untouched final test set to judge the completed search.



Forward selection follows measured increments

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Start with no features. Add lab, then age. Stop at lab+age if adding ward does not meet the improvement rule.

Candidate at current step	Illustrative validation score
Age alone / lab alone / ward alone	0.70 / 0.76 / 0.65
Lab+age / lab+ward	0.80 / 0.77
Lab+age+ward	0.79



Check. Greedy additions do not examine every combination. Pure interactions may fail to improve any one-feature candidate. Scores shown here illustrate the search, not actual experiments.

Backward elimination and RFE remove different candidates

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Start with all columns. Evaluate retrained subsets after removing one column and remove the least harmful candidate.

Check. Rankings can change after removal. Coefficients of differently scaled columns are not directly comparable. The desired remaining count needs validation selection.

Search	What is removed next?	What evidence selects the final subset?
Backward elimination	Compare validation results after candidate removals	A stated validation criterion.
RFE	Remove low-ranked features under the current fitted model	Validate the retained subset size.
Both	Refit after changes; interactions and substitutes can change rankings	Evaluate the complete search on held-out data.

Nested validation evaluates the whole search

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Concrete split.

Start with 100 development rows. A final test set is reserved separately.

Step	Rows used	Decision
Outer split	80 train, 20 held out	The 20 outcomes do not choose features
Inner comparison	Repeated splits within the 80	Compare complete candidate pipelines
Refit	All 80 outer-training rows	Fit the chosen pipeline from scratch
Outer score	The held-out 20	Evaluate the whole selection procedure

Rotate the outer split. Different folds may select different features.

Information boundary

Scaling, imputation, supervised selection and model fitting all belong inside the training side of each comparison.

L1 can select while fitting the model

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The squared-error term fits the outcomes. The L1 term charges for coefficient magnitude and can set some coefficients exactly to zero. The intercept is usually unpenalized.

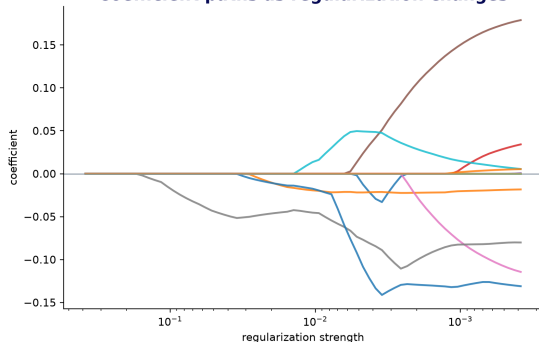
$$\min_{\beta_0, \beta} \frac{1}{2n} \|y - \beta_0 \mathbf{1} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{j=1}^p |\beta_j|.$$



Check. Scale features appropriately, choose λ in validation and compare with a dense baseline. Zero coefficients describe this model under this penalty, not universal irrelevance.

Follow coefficients as the L1 penalty weakens

Coefficient paths as regularization changes



The penalty weakens from left to right. Some coefficients leave zero as their cost decreases. Each curve represents a feature coefficient rather than a model score.



Coefficient size depends on units and substitutes

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

L1 charges coefficient size. A meaningful scale must be defined before comparing penalized coefficients.

$$\hat{y} = 2x \quad \hat{y} = 0.02x'$$

$$x' = 100x$$

Check. A sparse model can be useful and unstable in its chosen names at the same time. Inspect validation performance and selection frequencies rather than treating one ranking as permanent.

Observation	Why it can occur	How to check it
A large coefficient	A small input unit can require a large numerical coefficient	Compare appropriately scaled features.
One correlated feature gets zero	Its substitute carries similar predictive information	Inspect joint and grouped behavior.
The chosen variable changes after refitting	Several subsets may fit almost equally well	Compare stability and validation performance.

Lasso sets some linear coefficients to zero

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

These three models use the same advertising features. At this Lasso penalty, the newspaper coefficient is zero. Changing data, scaling or the penalty can change that result.

Term	Least squares	Ridge	Lasso
TV	3.983944	3.981524	3.921642
Radio	2.860230	2.858304	2.806374
Newspaper	0.038194	0.038295	0.000000
Intercept	13.969091	13.969282	13.972528



Check. A zero coefficient is not universal irrelevance. Correlated substitutes can change which feature remains.

Trees select conditions but importance needs validation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A tree chooses features while choosing splits. A fitted Iris tree may use petal length and petal width while leaving two sepal measurements unused.

Check. Impurity importance can prefer features with many possible cut points. Correlated substitutes can hide one another.

Evidence	Question answered	Question still open
Training splits	Which conditions this tree used	Will those rules work for new cases?
Impurity importance	How much fitted split improvement was attributed to a feature	Would intervening on it change the outcome?
Validation permutation	How predictions depend on that feature under this check	Are correlated alternatives masking the dependence?

Selection stability across repeated fits

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Illustrative repeated fits.

Run the same complete selection procedure on ten training resamples. Count how often each

	Feature	Selected runs	Frequency	Question to check
feature is retained.	Lab A	9 of 10	.9	Does the choice remain stable?
	Lab B	4 of 10	.4	Can it substitute for Lab A?
	Identifier	2 of 10	.2	Could chance splits explain its score?

A low frequency does not prove uselessness when correlated substitutes exist. A high frequency does not establish causality.

Compare prediction quality, measurement cost and availability at prediction time. Then save the chosen preprocessing and model as one pipeline.

A fitted pipeline preserves the complete calculation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Each fold learns preprocessing and the model from its training rows. A fitted selector belongs inside this pipeline.

```
num = make_pipeline(SimpleImputer(strategy="median"),
                    StandardScaler())
cat = make_pipeline(SimpleImputer(strategy="most_frequent"),
                    OneHotEncoder(handle_unknown="ignore"))
prep = ColumnTransformer([
    ("num", num, ["age", "lab"]),
    ("missing", MissingIndicator(features="all"), ["lab"]),
    ("cat", cat, ["ward"])]])
model = make_pipeline(pre, LogisticRegression(max_iter=1000))
scores = cross_val_score(model, X_train, y_train, cv=5)
```

One raw row becomes a vector with known coordinates

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The new row uses the saved imputation and scaling rules. Its lab value becomes the training median. A separate missingness indicator records that the lab was unmeasured, and ward C follows the unknown-category rule.

Worked calculation

$$\underbrace{(0.612, 0, 1, 0, 0)}_{\text{age z, lab z, lab missing, wardA, wardB}} .$$

Coordinate	Value	Reason
Age z	0.612	$(50 - 40)/16.33 \approx .612$.
Lab z	0	Impute training median 2; training lab mean is also 2.
Lab missing	1	The measurement was absent before imputation.
Ward A, Ward B	0, 0	New ward C is unknown to the saved A/B vocabulary.

Check. The new row did not refit the median, mean, scale or vocabulary.

Numerical results and information boundaries

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The training mean is 3 and the standard deviation is about 1.633. Standardizing 7 gives about 2.449. A z-score above 1 is valid. Fitted statistics and target information must still remain inside training.

$$z = \frac{7 - 3}{\sqrt{8/3}} \approx 2.449.$$

Question	Answer and reason
Encode targets and select features before splitting	Both steps leak; fit them inside each training fold
Predict at admission using days until discharge	Unavailable at prediction time; exclude it from inputs
XOR Both features have zero individual scores	Their joint relationship can still determine the label

A z-score above 1 is valid; its range is not restricted to $[0, 1]$.

Does an individual feature score settle feature usefulness?

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Question

In XOR, both inputs have zero marginal label information.
Should we discard both?

Evidence and reasoning

No: their pair determines the label. A univariate filter can miss interactions. Also, a zero Lasso coefficient describes one fitted dataset, scaling and penalty, rather than proving a feature is universally useless.

Check. Fit imputation, scaling, encoding and selection inside training folds before evaluating the resulting pipeline.