Data Science for Mathematicians
Exercises 4: Model Evaluation and Statistical Inference

Instructions

Answer all exercises completely. Show all working, justify your answers, and state any assumptions you make. For computational exercises, carry out all intermediate steps explicitly. For proof exercises, clearly identify which definitions and theorems you are applying.

Exercises

Exercise 1. Computing Regression Metrics

A linear regression model produces predictions $\hat{y}_i$ for $n = 6$ observations. The observed and predicted values are:

i	1	2	3	4	5	6
y_i	10	15	13	20	18	25
$\hat{y}_i$	11	14	14	19	20	23

(a) Compute the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

(b) Compute $\mathrm{SS}_{\mathrm{tot}}$ and $\mathrm{SS}_{\mathrm{res}}$.

(c) Compute the coefficient of determination R^2.

(d) Compute the Adjusted R^2, assuming the model uses $p = 1$ predictor.

Exercise 2. Confusion Matrix and Classification Metrics

A binary classifier is applied to $n = 250$ loan applications to predict default ($y = 1$) or non-default ($y = 0$). The results are:

$$\mathrm{TP} = 30, \quad \mathrm{FP} = 15, \quad \mathrm{FN} = 20, \quad \mathrm{TN} = 185.$$

(a) Verify that $\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN} = n$.

(b) Compute the Accuracy, Precision, Recall, and F_1-Score.

(c) Compute the False Positive Rate (FPR) and the Specificity.

(d) Compute the F_2-Score and the $\mathrm{F}_{0.5}$-Score. Which metric would you prioritize if the primary concern is ensuring that as many actual defaulters as possible are identified?

Exercise 3. Constructing an ROC Curve

A classifier outputs scores for $n = 8$ observations. The scores and true labels are given below, sorted by decreasing score:

Observation	1	2	3	4	5	6	7	8
Score $s(x_i)$	0.92	0.85	0.70	0.60	0.50	0.38	0.25	0.12
True label y_i	1	0	1	1	0	1	0	0

(a) Identify the number of actual positives P and actual negatives N.

(b) For each threshold τ (set at each score value and at > 0.92), compute the TP, FP, TPR, and FPR. Present your results in a table.

(c) Compute the AUC using the trapezoidal rule.

(d) Does this classifier perform better than random guessing? Justify your answer.

Exercise 4. Cross-Validation Computation

Consider a dataset with $n = 8$ observations:

x_i	1	2	3	4	5	6	7	8
y_i	2.1	4.3	5.8	8.2	10.1	12.5	13.9	16.0

We fit a linear model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ using OLS. We perform 4-fold CV with the following partition:

$$D_1 = \{(1, 2.1),\ (2, 4.3)\}, \quad D_2 = \{(3, 5.8),\ (4, 8.2)\},$$
$$D_3 = \{(5, 10.1),\ (6, 12.5)\}, \quad D_4 = \{(7, 13.9),\ (8, 16.0)\}.$$

For each fold, the OLS model trained on the remaining 6 points yields:

Fold	OLS equation
1	$\hat{y} = -0.37 + 2.05x$
2	$\hat{y} = -0.10 + 1.98x$
3	$\hat{y} = 0.22 + 1.93x$
4	$\hat{y} = -0.45 + 2.08x$

(a) For each fold, compute the predictions on the held-out observations and the fold MSE M_i.

(b) Compute the 4-fold cross-validation estimate CV_4.

(c) If LOOCV were used instead, how many models would need to be trained? What is the size of each training set?

Exercise 5. Deriving the Bias-Variance Decomposition for a Constant Estimator

Suppose the true data-generating process is $y = f(x) + \epsilon$ with $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. Consider the constant estimator $\hat{f}(x_0) = c$ for all x_0, where $c \in \mathbb{R}$ is a fixed constant.

(a) Compute $\mathrm{Bias}(\hat{f}(x_0))$ and $\mathrm{Var}(\hat{f}(x_0))$ for this estimator.

(b) Using the Bias-Variance Decomposition, write the expected prediction error $\mathbb{E}[(y_0 - \hat{f}(x_0))^2]$ in terms of $f(x_0)$, c, and σ^2.

(c) Find the value of c that minimizes the expected prediction error at x_0. What does this result tell you about the relationship between the optimal constant estimator and the true function?

(d) Now suppose c is not fixed but is instead the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ computed from a training set of n i.i.d. observations from the process. Compute $\mathbb{E}[\bar{y}]$ and $\mathrm{Var}(\bar{y})$, and hence find $\mathrm{Bias}(\bar{y})$ and $\mathrm{Var}(\bar{y})$ as an estimator of $f(x_0)$. Assume the training points x_i are distinct from x_0.

Exercise 6. R^2 and Added Predictors

Let R_1^2 denote the coefficient of determination for an OLS model with p_1 predictors and R_2^2 denote it for a model with $p_2 = p_1+1$ predictors (the second model includes all predictors of the first plus one additional predictor).

(a) Prove that $R_2^2 \geq R_1^2$. (Hint: use the fact that the OLS solution minimizes SS_{res} and that the parameter space of Model 1 is a subspace of that of Model 2.)

(b) Give a condition on the reduction in SS_{res} under which R^2_{adj} for the larger model exceeds that of the smaller model. Express your answer in terms of $SS_{res,1}$, n, and p_1.

(c) A colleague fits three nested linear models to $n = 50$ observations and reports:

Model	p	R^2
A	2	0.72
B	5	0.74
C	12	0.76

Compute R^2_{adj} for each model. Which model would you recommend and why?

Exercise 7. Properties of the F_1-Score

Let $P =$ Precision and $R =$ Recall, both in $(0, 1]$.

(a) Show that $F_1 = \frac{2PR}{P+R}$ can be rewritten as $\frac{1}{F_1} = \frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right)$, confirming it is the harmonic mean.

(b) Prove that $F_1 \leq \frac{P+R}{2}$ (the harmonic mean is at most the arithmetic mean), with equality if and only if $P = R$.

(c) A classifier on a test set of 200 samples achieves TP = 40, FP = 10, FN = 50, TN = 100. Compute P, R, and F_1. Then suppose we lower the decision threshold so that TP increases to 70 and FP increases to 40. Recompute P, R, and F_1, and comment on the precision-recall tradeoff.

Exercise 8. Bias and Variance of the CV Estimator

Consider a dataset of size n and a model class that requires at least m training observations to fit reliably (i.e., if the training set has fewer than m points, the fitted model is highly unstable).

(a) In k-fold CV, what is the size of each training set (in terms of n and k)? Derive a constraint on k that ensures each training set has at least m points.

(b) Explain why the LOOCV estimator ($k = n$) has low bias but potentially high variance. Relate your explanation to the overlap between training sets across folds.

(c) Suppose you observe the following 5-fold CV results for two models:

	M_1	M_2	M_3	M_4	M_5
Model X	3.1	3.3	3.0	3.2	3.4
Model Y	1.5	5.0	2.0	6.5	1.0

Compute CV_5 for both models. Then compute the sample standard deviation of the fold errors $\{M_i\}$ for each model. Which model would you choose and why?

Exercise 9. Choosing the Right Metric: Medical Screening

A hospital is evaluating two diagnostic models for detecting a rare form of cancer that affects 1 in 1000 patients. Both models are tested on 10,000 patients (10 with cancer, 9,990 without).

	TP	FP	FN	TN
Model A	9	50	1	9,940
Model B	7	5	3	9,985

(a) Compute the Accuracy, Precision, Recall, and F_1-Score for both models.

(b) Compute the F_2-Score for both models.

(c) A hospital administrator argues that Model B should be chosen because it has higher accuracy and higher precision. Write a mathematically informed response explaining why Model A may be the better choice for a screening test. Reference the concepts of Type I and Type II errors and the relative costs in this medical context.

(d) What additional information or metric would you want to examine before making a final recommendation?

Exercise 10. Diagnosing Overfitting with Cross-Validation

A data scientist fits polynomial regression models of degrees $d = 1, 2, 3, 5, 8$ to a dataset with $n = 50$ observations. She reports the training MSE and the 10-fold CV MSE for each:

Degree d	Training MSE	CV MSE
1	8.50	9.20
2	3.10	3.85
3	2.80	3.40
5	1.20	5.70
8	0.15	12.40

(a) Which model appears to be underfitting? Which model(s) appear to be overfitting? Justify your answers using the bias-variance tradeoff.

(b) Explain why the training MSE decreases monotonically with d but the CV MSE does not.

(c) Which degree would you recommend? Justify your choice.

(d) Sketch a qualitative plot (or describe in precise mathematical language) showing how you expect the squared bias, variance, and total expected error to behave as a function of d for this scenario.

Exercise 11. Stratified Cross-Validation Design

You have a binary classification dataset with $n = 300$ observations: 240 belong to class 0 and 60 belong to class 1. You plan to use 5-fold cross-validation.

(a) If folds are created by unstratified random sampling, what is the expected number of class 1 observations per fold? In the worst case, a fold might receive very few or zero class 1 observations. Explain why this is problematic for estimating metrics like Recall.

(b) In stratified 5-fold CV, how many class 0 and class 1 observations should each fold contain?

(c) Describe an algorithm (in pseudocode or precise mathematical steps) for constructing a stratified partition of the dataset into k folds.

(d) After completing stratified 5-fold CV, how would you report the final performance? Discuss whether reporting only CV_5 is sufficient, or whether additional information (such as per-fold metrics or confidence intervals) should be included.

Exercise 12. Model Selection Pipeline

You are building a classifier for fraud detection on a dataset with $n = 5000$ transactions, of which 2% are fraudulent. You have three candidate models: Logistic Regression, k-Nearest Neighbors (k-NN), and a Decision Tree. Each model has hyperparameters that need tuning.

(a) Explain why accuracy is an inappropriate evaluation metric for this task. What metric(s) would you use instead, and why?

(b) Outline a complete model selection and evaluation pipeline using cross-validation. Your pipeline should include: (i) a method for hyperparameter tuning, (ii) a method for comparing models, and (iii) a final evaluation step. Be precise about which data is used at each stage.

(c) A colleague proposes the following approach: "Tune each model's hyperparameters using 5-fold CV, then evaluate the best configuration of each model on the same 5-fold CV to compare them, then report the best CV score as the expected performance." Identify the flaw in this approach and explain how it could lead to an optimistic estimate of generalization error.

Solutions

Solution 1. Computing Regression Metrics

(a) First, compute the residuals $e_i = y_i - \hat{y}_i$:

$$e_1 = -1, \quad e_2 = 1, \quad e_3 = -1, \quad e_4 = 1, \quad e_5 = -2, \quad e_6 = 2.$$

The squared residuals are $e_i^2 = 1, 1, 1, 1, 4, 4$. Therefore:

$$\text{MSE} = \frac{1}{6}(1 + 1 + 1 + 1 + 4 + 4) = \frac{12}{6} = 2.0.$$

$$\text{RMSE} = \sqrt{2.0} = \sqrt{2} \approx 1.414.$$

(b) The sample mean is $\bar{y} = \frac{10+15+13+20+18+25}{6} = \frac{101}{6} \approx 16.833$.
The deviations from the mean are:

$$(y_i - \bar{y})^2 : \left(\frac{-41}{6}\right)^2, \left(\frac{-11}{6}\right)^2, \left(\frac{-23}{6}\right)^2, \left(\frac{19}{6}\right)^2, \left(\frac{7}{6}\right)^2, \left(\frac{49}{6}\right)^2$$
$$= \frac{1681}{36}, \frac{121}{36}, \frac{529}{36}, \frac{361}{36}, \frac{49}{36}, \frac{2401}{36}.$$

$$\text{SS}_{\text{tot}} = \frac{1681 + 121 + 529 + 361 + 49 + 2401}{36} = \frac{5142}{36} = \frac{857}{6} \approx 142.833.$$

$$\text{SS}_{\text{res}} = \sum_{i=1}^{6} (y_i - \hat{y}_i)^2 = 12.$$

(c) Using the definition:

$$\text{R}^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} = 1 - \frac{12}{857/6} = 1 - \frac{72}{857} = \frac{785}{857} \approx 0.916.$$

(d) With $n = 6$ and $p = 1$:

$$\text{R}^2_{\text{adj}} = 1 - \frac{(1 - \text{R}^2)(n-1)}{n-p-1} = 1 - \frac{(1 - 0.916)(6-1)}{6-1-1} = 1 - \frac{0.084 \times 5}{4} = 1 - 0.105 = 0.895.$$

More precisely, $\text{R}^2_{\text{adj}} = 1 - \frac{(72/857)\cdot 5}{4} = 1 - \frac{360}{3428} = 1 - \frac{90}{857} = \frac{767}{857} \approx 0.895$.

Solution 2. Confusion Matrix and Classification Metrics

(a) $\text{TP} + \text{FP} + \text{FN} + \text{TN} = 30 + 15 + 20 + 185 = 250 = n$. ✓

(b) **Accuracy:**

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{n} = \frac{30 + 185}{250} = \frac{215}{250} = 0.86.$$

Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{30}{30 + 15} = \frac{30}{45} = \frac{2}{3} \approx 0.667.$$

Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{30}{30 + 20} = \frac{30}{50} = 0.60.$$

F_1**-Score:**

$$\text{F}_1 = \frac{2 \cdot \frac{2}{3} \cdot 0.60}{\frac{2}{3} + 0.60} = \frac{2 \cdot \frac{2}{3} \cdot \frac{3}{5}}{\frac{2}{3} + \frac{3}{5}} = \frac{\frac{4}{5}}{\frac{19}{15}} = \frac{4}{5} \cdot \frac{15}{19} = \frac{60}{95} = \frac{12}{19} \approx 0.632.$$

(c) **FPR:**

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{15}{15 + 185} = \frac{15}{200} = 0.075.$$

Specificity:

$$\text{Specificity} = 1 - \text{FPR} = 1 - 0.075 = 0.925.$$

(d) With $P = 2/3$ and $R = 3/5$:
F_2**-Score** ($\beta = 2$, emphasizes recall):

$$\text{F}_2 = \frac{(1+4) \cdot \frac{2}{3} \cdot \frac{3}{5}}{4 \cdot \frac{2}{3} + \frac{3}{5}} = \frac{5 \cdot \frac{2}{5}}{\frac{8}{3} + \frac{3}{5}} = \frac{2}{\frac{49}{15}} = \frac{30}{49} \approx 0.612.$$

$\text{F}_{0.5}$**-Score** ($\beta = 0.5$, emphasizes precision):

$$\text{F}_{0.5} = \frac{(1+0.25) \cdot \frac{2}{3} \cdot \frac{3}{5}}{0.25 \cdot \frac{2}{3} + \frac{3}{5}} = \frac{1.25 \cdot \frac{2}{5}}{\frac{1}{6} + \frac{3}{5}} = \frac{\frac{1}{2}}{\frac{23}{30}} = \frac{15}{23} \approx 0.652.$$

Since the primary concern is identifying actual defaulters (minimizing false negatives), we should prioritize **Recall** and use the F_2-Score, which weights recall more heavily than precision.

Solution 3. Constructing an ROC Curve

(a) There are $P = 4$ actual positives (observations 1, 3, 4, 6) and $N = 4$ actual negatives (observations 2, 5, 7, 8).

(b) Sweeping the threshold from high to low:

Threshold τ	TP	FP	TPR	FPR
> 0.92 (all negative)	0	0	0.00	0.00
0.92	1	0	0.25	0.00
0.85	1	1	0.25	0.25
0.70	2	1	0.50	0.25
0.60	3	1	0.75	0.25
0.50	3	2	0.75	0.50
0.38	4	2	1.00	0.50
0.25	4	3	1.00	0.75
0.12	4	4	1.00	1.00

(c) Using the trapezoidal rule, we sum the area under each horizontal/vertical step. The FPR changes at the following transitions (listing pairs (FPR, TPR)):
$(0, 0) \to (0, 0.25)$: vertical, no area.
$(0, 0.25) \to (0.25, 0.25)$: area $= 0.25 \times 0.25 = 0.0625$.
$(0.25, 0.25) \to (0.25, 0.75)$: vertical, no area.
$(0.25, 0.75) \to (0.50, 0.75)$: area $= 0.75 \times 0.25 = 0.1875$.
$(0.50, 0.75) \to (0.50, 1.00)$: vertical, no area.
$(0.50, 1.00) \to (0.75, 1.00)$: area $= 1.00 \times 0.25 = 0.25$.
$(0.75, 1.00) \to (1.00, 1.00)$: area $= 1.00 \times 0.25 = 0.25$.

$$\text{AUC} = 0.0625 + 0.1875 + 0.25 + 0.25 = 0.75.$$

(d) Yes. The AUC $= 0.75 > 0.5$, which means the classifier has significantly better discriminative ability than random guessing (AUC $= 0.5$). The probabilistic interpretation is that there is a 75% chance that the classifier assigns a higher score to a randomly chosen positive instance than to a randomly chosen negative instance.

Solution 4. Cross-Validation Computation

(a) **Fold 1:** Held-out points $(1, 2.1)$ and $(2, 4.3)$. Using $\hat{y} = -0.37 + 2.05x$:

$$\hat{y}(1) = -0.37 + 2.05 = 1.68, \qquad \hat{y}(2) = -0.37 + 4.10 = 3.73.$$

$$M_1 = \frac{1}{2}\left[(2.1 - 1.68)^2 + (4.3 - 3.73)^2\right] = \frac{0.1764 + 0.3249}{2} = \frac{0.5013}{2} \approx 0.251.$$

Fold 2: Held-out points $(3, 5.8)$ and $(4, 8.2)$. Using $\hat{y} = -0.10 + 1.98x$:

$$\hat{y}(3) = -0.10 + 5.94 = 5.84, \qquad \hat{y}(4) = -0.10 + 7.92 = 7.82.$$

$$M_2 = \frac{1}{2}\left[(5.8 - 5.84)^2 + (8.2 - 7.82)^2\right] = \frac{0.0016 + 0.1444}{2} = \frac{0.1460}{2} \approx 0.073.$$

Fold 3: Held-out points $(5, 10.1)$ and $(6, 12.5)$. Using $\hat{y} = 0.22 + 1.93x$:

$$\hat{y}(5) = 0.22 + 9.65 = 9.87, \qquad \hat{y}(6) = 0.22 + 11.58 = 11.80.$$

$$M_3 = \frac{1}{2}\left[(10.1 - 9.87)^2 + (12.5 - 11.80)^2\right] = \frac{0.0529 + 0.49}{2} = \frac{0.5429}{2} \approx 0.271.$$

Fold 4: Held-out points $(7, 13.9)$ and $(8, 16.0)$. Using $\hat{y} = -0.45 + 2.08x$:

$$\hat{y}(7) = -0.45 + 14.56 = 14.11, \qquad \hat{y}(8) = -0.45 + 16.64 = 16.19.$$

$$M_4 = \frac{1}{2}\left[(13.9 - 14.11)^2 + (16.0 - 16.19)^2\right] = \frac{0.0441 + 0.0361}{2} = \frac{0.0802}{2} \approx 0.040.$$

(b) The 4-fold CV estimate is:

$$\mathrm{CV}_4 = \frac{1}{4}(M_1 + M_2 + M_3 + M_4) = \frac{0.251 + 0.073 + 0.271 + 0.040}{4} = \frac{0.635}{4} \approx 0.159.$$

(c) With LOOCV ($k = n = 8$), we would need to train 8 separate models. Each training set would have size $n - 1 = 7$.

Solution 5. Deriving the Bias-Variance Decomposition for a Constant Estimator

(a) Since $\hat{f}(x_0) = c$ is a fixed constant (not dependent on the training data D), we have:

$$\mathbb{E}_D[\hat{f}(x_0)] = c.$$

Therefore:

$$\mathrm{Bias}(\hat{f}(x_0)) = \mathbb{E}_D[\hat{f}(x_0)] - f(x_0) = c - f(x_0).$$

$$\mathrm{Var}(\hat{f}(x_0)) = \mathbb{E}_D\left[(\hat{f}(x_0) - \mathbb{E}_D[\hat{f}(x_0)])^2\right] = \mathbb{E}_D\left[(c - c)^2\right] = 0.$$

(b) By the Bias-Variance Decomposition (Theorem 1):

$$\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = \mathrm{Bias}^2 + \mathrm{Var} + \sigma^2 = (c - f(x_0))^2 + 0 + \sigma^2 = (c - f(x_0))^2 + \sigma^2.$$

(c) To minimize over c, we take the derivative and set it to zero:

$$\frac{\mathrm{d}}{\mathrm{d}c}\left[(c - f(x_0))^2 + \sigma^2\right] = 2(c - f(x_0)) = 0 \implies c^* = f(x_0).$$

The optimal constant estimator at x_0 is the true function value $f(x_0)$. This makes intuitive sense: the best fixed prediction is the true signal, yielding zero bias. The remaining error is purely the irreducible noise σ^2.

(d) Now $\hat{f}(x_0) = \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ where $y_i = f(x_i) + \epsilon_i$.

$$\mathbb{E}[\bar{y}] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[y_i] = \frac{1}{n}\sum_{i=1}^n f(x_i) = \bar{f},$$

where $\bar{f} = \frac{1}{n}\sum_{i=1}^n f(x_i)$ is the average of the true function values at the training points.

Since the ϵ_i are independent with $\mathrm{Var}(\epsilon_i) = \sigma^2$:

$$\mathrm{Var}(\bar{y}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n (f(x_i) + \epsilon_i)\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\epsilon_i) = \frac{\sigma^2}{n}.$$

As an estimator of $f(x_0)$:

$$\mathrm{Bias}(\bar{y}) = \mathbb{E}[\bar{y}] - f(x_0) = \bar{f} - f(x_0).$$

The bias is nonzero unless $f(x_0)$ happens to equal the average of f over the training inputs. The variance σ^2/n decreases with sample size, illustrating that more data reduces the variance component of error.

Solution 6. R^2 **and Added Predictors**

(a) Model 1 minimizes $\mathrm{SS}_{\mathrm{res}}$ over coefficients $(\beta_0, \beta_1, \ldots, \beta_{p_1})$. Model 2 minimizes $\mathrm{SS}_{\mathrm{res}}$ over $(\beta_0, \beta_1, \ldots, \beta_{p_1}, \beta_{p_1+1})$. The parameter space of Model 1 is a subspace of Model 2's parameter space (obtained by setting $\beta_{p_1+1} = 0$).

Since OLS finds the global minimizer of $\mathrm{SS}_{\mathrm{res}}$ over the parameter space, the minimum over a larger set is at most the minimum over a subset:

$$\mathrm{SS}_{\mathrm{res},2} \leq \mathrm{SS}_{\mathrm{res},1}.$$

Since $\mathrm{SS}_{\mathrm{tot}}$ is independent of the model:

$$\mathrm{R}_2^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res},2}}{\mathrm{SS}_{\mathrm{tot}}} \geq 1 - \frac{\mathrm{SS}_{\mathrm{res},1}}{\mathrm{SS}_{\mathrm{tot}}} = \mathrm{R}_1^2.$$

(b) The Adjusted R^2 for Model 1 is:

$$\mathrm{R}^2_{\mathrm{adj},1} = 1 - \frac{\mathrm{SS}_{\mathrm{res},1}/(n - p_1 - 1)}{\mathrm{SS}_{\mathrm{tot}}/(n-1)}.$$

For Model 2 (with $p_2 = p_1 + 1$):

$$\mathrm{R}^2_{\mathrm{adj},2} = 1 - \frac{\mathrm{SS}_{\mathrm{res},2}/(n - p_1 - 2)}{\mathrm{SS}_{\mathrm{tot}}/(n-1)}.$$

For $\mathrm{R}^2_{\mathrm{adj},2} > \mathrm{R}^2_{\mathrm{adj},1}$, we need:

$$\frac{\mathrm{SS}_{\mathrm{res},2}}{n - p_1 - 2} < \frac{\mathrm{SS}_{\mathrm{res},1}}{n - p_1 - 1}.$$

That is, the new predictor must reduce $\mathrm{SS}_{\mathrm{res}}$ by enough so that the residual mean square (mean $\mathrm{SS}_{\mathrm{res}}$ per degree of freedom) decreases. Rearranging:

$$\mathrm{SS}_{\mathrm{res},2} < \mathrm{SS}_{\mathrm{res},1} \cdot \frac{n - p_1 - 2}{n - p_1 - 1}.$$

Equivalently, the reduction $\Delta = \mathrm{SS}_{\mathrm{res},1} - \mathrm{SS}_{\mathrm{res},2}$ must satisfy:

$$\Delta > \frac{\mathrm{SS}_{\mathrm{res},1}}{n - p_1 - 1}.$$

(c) For $n = 50$:
Model A ($p = 2$, $\mathrm{R}^2 = 0.72$):

$$\mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{(1 - 0.72)(49)}{47} = 1 - \frac{0.28 \times 49}{47} = 1 - \frac{13.72}{47} = 1 - 0.2919 = 0.708.$$

Model B ($p = 5$, $\mathrm{R}^2 = 0.74$):

$$\mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{(1 - 0.74)(49)}{44} = 1 - \frac{0.26 \times 49}{44} = 1 - \frac{12.74}{44} = 1 - 0.2895 = 0.710.$$

Model C ($p = 12$, $\mathrm{R}^2 = 0.76$):

$$\mathrm{R}^2_{\mathrm{adj}} = 1 - \frac{(1 - 0.76)(49)}{37} = 1 - \frac{0.24 \times 49}{37} = 1 - \frac{11.76}{37} = 1 - 0.3178 = 0.682.$$

Although Model C has the highest R^2, it has the *lowest* $\mathrm{R}^2_{\mathrm{adj}}$. Model B has the highest $\mathrm{R}^2_{\mathrm{adj}}$ at 0.710, but it is barely higher than Model A's 0.708. Given the principle of parsimony and the negligible improvement, **Model A** (with only 2 predictors) is arguably the best choice. At a minimum, Model C should be rejected as its 10 additional predictors beyond Model A actually hurt the adjusted metric.

Solution 7. Properties of the F_1-Score

(a) Starting from $\mathrm{F}_1 = \frac{2PR}{P+R}$, take the reciprocal:

$$\frac{1}{\mathrm{F}_1} = \frac{P + R}{2PR} = \frac{P}{2PR} + \frac{R}{2PR} = \frac{1}{2R} + \frac{1}{2P} = \frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right).$$

This is the definition of the harmonic mean: $\mathrm{F}_1 = H(P, R)$.

(b) We wish to prove $\frac{2PR}{P+R} \leq \frac{P+R}{2}$. Cross-multiplying (all terms are positive):

$$4PR \leq (P + R)^2 = P^2 + 2PR + R^2.$$

This simplifies to:

$$0 \leq P^2 - 2PR + R^2 = (P - R)^2.$$

Since $(P - R)^2 \geq 0$ always holds, the inequality is proven. Equality holds if and only if $(P - R)^2 = 0$, i.e., $P = R$.

(c) **Before lowering threshold:** TP = 40, FP = 10, FN = 50, TN = 100.

$$P = \frac{40}{40+10} = \frac{40}{50} = 0.80, \qquad R = \frac{40}{40+50} = \frac{40}{90} \approx 0.444.$$

$$\mathrm{F}_1 = \frac{2 \times 0.80 \times 0.444}{0.80 + 0.444} = \frac{0.711}{1.244} \approx 0.571.$$

After lowering threshold: TP = 70, FP = 40, FN = 90 − 70 = 20, TN = 100 − 30 = 70.
(Note: total positives = TP + FN = 90, total negatives = FP + TN = 110, unchanged.)

$$P = \frac{70}{70+40} = \frac{70}{110} = \frac{7}{11} \approx 0.636, \qquad R = \frac{70}{70+20} = \frac{70}{90} = \frac{7}{9} \approx 0.778.$$

$$\mathrm{F}_1 = \frac{2 \times \frac{7}{11} \times \frac{7}{9}}{\frac{7}{11} + \frac{7}{9}} = \frac{2 \times \frac{49}{99}}{\frac{63+77}{99}} = \frac{\frac{98}{99}}{\frac{140}{99}} = \frac{98}{140} = 0.70.$$

Comment: Lowering the threshold increased Recall from 0.444 to 0.778 (a 75% relative increase) but decreased Precision from 0.80 to 0.636. This is the precision-recall tradeoff in action: a more liberal threshold captures more true positives but also generates more false positives. The F_1-score increased from 0.571 to 0.70, suggesting that for this particular model the original threshold was too conservative.

Solution 8. Bias and Variance of the CV Estimator

(a) In k-fold CV, each training set has size $\frac{k-1}{k} \cdot n$. For this to be at least m:

$$\frac{k-1}{k} \cdot n \geq m \implies k - 1 \geq \frac{mk}{n} \implies k\left(1 - \frac{m}{n}\right) \geq 1 \implies k \geq \frac{n}{n-m}.$$

Since k must also satisfy $k \leq n$, the constraint is $\frac{n}{n-m} \leq k \leq n$ (and $n > m$).

(b) With LOOCV ($k = n$), each training set has size $n - 1$, which is nearly the size of the full dataset. This means the model trained in each fold closely approximates the model trained on all n points. Consequently, each M_i estimates the error of a model trained on nearly all the data, so the bias of the CV estimate (relative to the error of the full-data model) is approximately zero.

However, any two training sets share $n - 2$ of their $n - 1$ observations. This extreme overlap makes the n fitted models $\hat{f}_1, \ldots, \hat{f}_n$ highly correlated, and so are their errors $e_1, \ldots, e_n$. The variance of the average $\mathrm{CV}_n = \frac{1}{n}\sum e_i$ depends not only on $\frac{1}{n^2}\sum \mathrm{Var}(e_i)$ but also on the covariance terms $\frac{1}{n^2}\sum_{i \neq j} \mathrm{Cov}(e_i, e_j)$. With highly positive covariances, the averaging provides little variance reduction, leading to a potentially high-variance estimate.

(c) For Model X:

$$\mathrm{CV}_5^X = \frac{3.1 + 3.3 + 3.0 + 3.2 + 3.4}{5} = \frac{16.0}{5} = 3.20.$$

For Model Y:

$$\mathrm{CV}_5^Y = \frac{1.5 + 5.0 + 2.0 + 6.5 + 1.0}{5} = \frac{16.0}{5} = 3.20.$$

Both models have the same average CV error of 3.20.

Sample standard deviation (with $k-1=4$ in the denominator):

For Model X: $\bar{M}=3.20$.

$$s_X=\sqrt{\frac{(3.1-3.2)^2+(3.3-3.2)^2+(3.0-3.2)^2+(3.2-3.2)^2+(3.4-3.2)^2}{4}}$$
$$=\sqrt{\frac{0.01+0.01+0.04+0+0.04}{4}}=\sqrt{\frac{0.10}{4}}\approx 0.158.$$

For Model Y: $\bar{M}=3.20$.

$$s_Y=\sqrt{\frac{(1.5-3.2)^2+(5.0-3.2)^2+(2.0-3.2)^2+(6.5-3.2)^2+(1.0-3.2)^2}{4}}$$
$$=\sqrt{\frac{2.89+3.24+1.44+10.89+4.84}{4}}=\sqrt{\frac{23.30}{4}}\approx 2.414.$$

Recommendation: Although both models have identical CV_5, Model X should be preferred because its fold errors are highly consistent ($s_X\approx 0.16$), indicating stable generalization performance. Model Y's large variation ($s_Y\approx 2.41$) suggests it is sensitive to the particular data split, a hallmark of high-variance models that may be overfitting. A model with stable cross-validation performance is more trustworthy for deployment.

Solution 9. Choosing the Right Metric: Medical Screening

(a) **Model A:**

$$\text{Accuracy}=\frac{9+9940}{10000}=\frac{9949}{10000}=0.9949,$$
$$\text{Precision}=\frac{9}{9+50}=\frac{9}{59}\approx 0.153,$$
$$\text{Recall}=\frac{9}{9+1}=\frac{9}{10}=0.90,$$
$$\mathrm{F}_1=\frac{2\times 0.153\times 0.90}{0.153+0.90}=\frac{0.2754}{1.053}\approx 0.262.$$

Model B:

$$\text{Accuracy}=\frac{7+9985}{10000}=\frac{9992}{10000}=0.9992,$$
$$\text{Precision}=\frac{7}{7+5}=\frac{7}{12}\approx 0.583,$$
$$\text{Recall}=\frac{7}{7+3}=\frac{7}{10}=0.70,$$
$$\mathrm{F}_1=\frac{2\times 0.583\times 0.70}{0.583+0.70}=\frac{0.8162}{1.283}\approx 0.636.$$

(b) **F_2-Score for Model A:**

$$\mathrm{F}_2=\frac{5\times 0.153\times 0.90}{4\times 0.153+0.90}=\frac{0.6885}{0.612+0.90}=\frac{0.6885}{1.512}\approx 0.455.$$

F_2-Score for Model B:

$$\mathrm{F}_2=\frac{5\times 0.583\times 0.70}{4\times 0.583+0.70}=\frac{2.0405}{2.332+0.70}=\frac{2.0405}{3.032}\approx 0.673.$$

(c) While Model B has higher accuracy (99.92% vs. 99.49%) and higher precision (58.3% vs. 15.3%), these metrics are misleading for a cancer screening application.

In a screening test for a life-threatening disease, a **false negative** (Type II error) means a cancer patient is told they are healthy and does not receive timely treatment. This can be fatal. A **false positive** (Type I error) means a healthy patient undergoes additional testing—inconvenient and anxiety-inducing, but not dangerous.

The asymmetry in costs is extreme: missing cancer $\gg$ ordering an extra test. Therefore, **Recall** (sensitivity) is the most critical metric. Model A has recall of 0.90, detecting 9 out of 10 cancer patients, while Model B has recall of only 0.70, missing 3 out of 10 cancer patients. Model A misses 1 cancer case; Model B misses 3—three times as many.

The high accuracy of both models is an artifact of the severe class imbalance (only 0.1% prevalence). A trivial model predicting "no cancer" for everyone would achieve 99.9% accuracy while being completely useless.

For a **screening** context where the goal is to flag potential cases for further diagnostic testing, **Model A is the better choice** despite its lower precision, because it saves more lives by identifying more true cases.

(d) Before making a final recommendation, one should examine:

- The full **ROC curve** and **Precision-Recall curve** for both models, to understand performance across all thresholds (not just the default).
- The **AUC** for a threshold-independent comparison of discriminative ability.
- Whether the decision threshold can be adjusted—Model B at a lower threshold might achieve recall comparable to Model A.
- The **cost** of follow-up diagnostic procedures to quantify the economic impact of false positives.
- Confidence intervals on the metrics, given the very small number of positive cases ($n = 10$), which makes all positive-class metrics highly uncertain.

Solution 10. Diagnosing Overfitting with Cross-Validation

(a) **Underfitting:** The $d = 1$ (linear) model has the highest CV MSE (9.20) and also a relatively high training MSE (8.50). Both errors being high indicates the model lacks the flexibility to capture the underlying pattern. This is characteristic of high bias / low variance, i.e., underfitting.

Overfitting: The $d = 5$ and $d = 8$ models show a large gap between training MSE and CV MSE. For $d = 8$, the training MSE is extremely low (0.15) while the CV MSE is very high (12.40). The model memorizes the training data (low training error) but fails to generalize (high test error). This is characteristic of low bias / high variance, i.e., overfitting. The $d = 5$ model (Training MSE = 1.20, CV MSE = 5.70) also shows overfitting, though less severe.

(b) The training MSE decreases monotonically because adding more polynomial terms strictly expands the hypothesis space. By the same argument as for R^2, the OLS fit on the training data can only improve (or stay the same) with more parameters.

The CV MSE does not decrease monotonically because it measures performance on *held-out* data. Beyond the optimal complexity, additional parameters fit the noise in the training folds rather than the signal. This noise-fitting does not transfer to the validation fold, causing the CV error to increase. The CV MSE reflects the sum $\text{Bias}^2 + \text{Var} + \sigma^2$, where the variance term grows with model complexity faster than the bias term shrinks.

(c) The recommended degree is $d = 3$, which achieves the lowest CV MSE of 3.40. It offers the best tradeoff between bias (reduced relative to $d = 1, 2$) and variance (controlled relative to $d = 5, 8$). Alternatively, $d = 2$ with CV MSE $= 3.85$ is also a strong candidate; if one applies the "one standard error rule" (selecting the simplest model within one SE of the minimum), $d = 2$ might be preferred for parsimony. Without standard error information, $d = 3$ is the best choice based on the CV MSE.

(d) As d increases:

- **Squared Bias** decreases monotonically, starting high (the linear model cannot capture curvature) and approaching zero (high-degree polynomials can approximate any smooth function).
- **Variance** increases monotonically, starting near zero (a linear model is stable across datasets) and growing rapidly (high-degree polynomials are extremely sensitive to the training data).
- **Irreducible error** σ^2 is constant, forming a horizontal baseline.
- **Total expected error** $= \text{Bias}^2 + \text{Var} + \sigma^2$ is U-shaped: it starts high (dominated by bias), decreases to a minimum around $d = 3$, then increases again (dominated by variance).

The minimum of the U-shaped total error curve corresponds to the optimal model complexity.

Solution 11. Stratified Cross-Validation Design

(a) With $n = 300$ and $k = 5$, each fold has $300/5 = 60$ observations. The expected number of class 1 observations per fold is $60/300 \times 60 = 12$ (or equivalently, $60 \times (60/300) = 12$).

However, with unstratified random splitting, the actual count per fold is a hypergeometric random variable. In the worst case, a fold might receive 0 or very few class 1 instances. This is problematic because:

- Recall $= \text{TP}/(\text{TP} + \text{FN})$ requires actual positive instances in the validation fold to be meaningful. With 0 positives, recall is undefined.
- Even with a few positives (say 1 or 2), the metric estimate is dominated by sampling noise and is unreliable.
- The training set would then be missing a proportional share of positive examples, potentially degrading the model's ability to learn the minority class pattern.

(b) In stratified 5-fold CV, each fold should contain:

$$\text{Class 0 per fold} = \frac{240}{5} = 48, \qquad \text{Class 1 per fold} = \frac{60}{5} = 12.$$

Each fold has $48 + 12 = 60$ observations, preserving the original 80%/20% class proportion.

(c) Algorithm for stratified k-fold partition:

i. Separate the dataset into class-specific subsets: $D_0 = \{i : y_i = 0\}$ and $D_1 = \{i : y_i = 1\}$.

ii. Randomly shuffle D_0 and D_1 independently.

iii. Partition D_0 into k groups of size $\lfloor |D_0|/k \rfloor$ (distributing the $|D_0|$ mod k remainder observations one each to the first folds).

iv. Partition D_1 into k groups of size $\lfloor |D_1|/k \rfloor$ (distributing the remainder similarly).

v. Form fold j by concatenating the j-th group from D_0 and the j-th group from D_1, for $j = 1, \ldots, k$.

This guarantees that each fold has approximately the same class proportions as the full dataset.

(d) Reporting only CV_5 (the mean) is insufficient. A thorough report should include:

- The mean CV_5 and the **standard deviation** (or standard error) of the per-fold metrics $\{M_1, \ldots, M_5\}$, to quantify the stability of the estimate.
- The **per-fold metrics** $M_1, \ldots, M_5$, so the reader can assess whether one fold is an outlier.
- A **confidence interval** for the generalization error, e.g., $\mathrm{CV}_5 \pm t_{0.025,4} \cdot s/\sqrt{5}$, acknowledging that the fold errors are not independent (due to overlapping training sets) and thus the interval may be approximate.
- For classification, reporting multiple metrics (Precision, Recall, F_1) per fold, not just a single aggregate, to understand which aspects of performance vary.

Solution 12. Model Selection Pipeline

(a) With only 2% of transactions being fraudulent, the dataset is highly imbalanced. A trivial classifier predicting "not fraud" for every transaction achieves 98% accuracy, yet detects zero fraud cases.

More appropriate metrics include:

- F_1**-Score** or F_2**-Score**: These balance precision and recall, with F_2 being especially appropriate if missing a fraudulent transaction (false negative) is more costly than flagging a legitimate one (false positive).
- **Precision and Recall separately**: To understand the nature of errors.
- **AUC-ROC** or **AUC-PR** (Area Under the Precision-Recall Curve): These provide threshold-independent assessments. AUC-PR is preferred over AUC-ROC for severely imbalanced problems, as the ROC curve can give an overly optimistic picture.

(b) A rigorous pipeline:

i. **Initial split:** Partition D into a development set D_{dev} (e.g., 80%) and a held-out test set D_{test} (20%), using stratified sampling to preserve the 2% fraud rate in both sets.

ii. **Hyperparameter tuning (inner CV):** For each candidate model, perform stratified 5-fold CV on D_{dev} for each hyperparameter configuration. Select the hyperparameters that maximize the chosen metric (e.g., F_2-Score) averaged over the 5 folds.

iii. **Model comparison (outer CV or direct):** Using the best hyperparameters for each model, either (a) perform another round of stratified

5-fold CV on D_{dev} (nested CV), or (b) compare the inner CV scores from step 2.

iv. **Final evaluation:** Select the best model, retrain it on the *entire* D_{dev} with the chosen hyperparameters, and evaluate once on D_{test}. This gives an unbiased estimate of generalization error. Report this as the expected performance.

(c) The flaw is that the colleague uses the **same data** for both hyperparameter tuning and model comparison. When the hyperparameters are selected via 5-fold CV to maximize performance, the resulting CV score is biased upward because the hyperparameters were optimized to perform well on exactly these folds. This is a form of **data leakage**: the validation folds indirectly influenced the model configuration through the tuning process.

Reporting this optimistically biased CV score as the expected performance will overestimate how well the model generalizes to truly unseen data. The correct approach uses **nested cross-validation** (an inner loop for tuning and an outer loop for evaluation) or a separate held-out test set that is never used during any model selection or tuning step.