

SDS 336 - Midterm Examination

80 minutes 40 questions 2 points each 80 points total

General instructions for Parts I and II. One sheet of letter paper, printed on both sides, is permitted. Select one best answer (A–D) for each question.

The exam has seven settings: four synthetic settings in Part I and three settings based on real data in Part II. A setting is a shared problem context for several consecutive questions; answer each question independently using the information provided. In Part II, all displayed code has already been run; use the code and output as evidence.

Part I: Conceptual questions

Questions 1–20 use Settings 1–4, which are synthetic examples created for this exam.

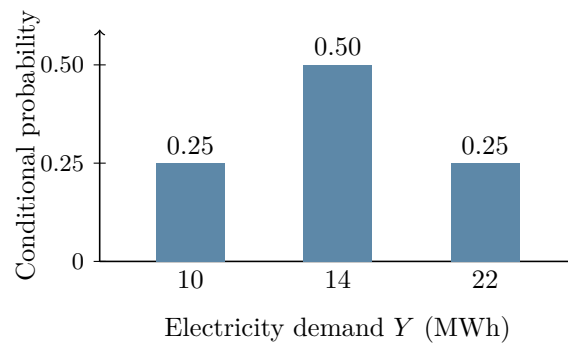
Setting 1: Electricity demand regression

An electric utility studies and predicts electricity demand using weather, calendar, and neighborhood information. The questions below examine optimal prediction, nonlinear features, model flexibility, and local smoothing.

Question 1

2 points

For neighborhoods with predictor profile $X = x_0$, historical evidence gives the conditional distribution below.



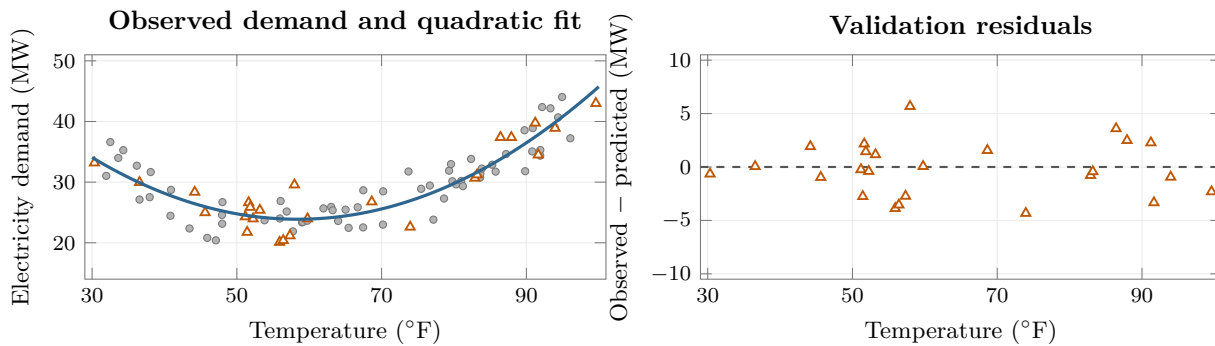
The utility evaluates predictions using squared error. What should it predict for a neighborhood with $X = x_0$?

- A. 22 MWh
- B. 15 MWh
- C. 14 MWh
- D. 10 MWh

Question 2

2 points

The utility adds $(T - 65)^2$ as a new predictor, where T is temperature, and refits the model. The resulting quadratic fit in temperature and its validation residuals are shown below.



Model	Training RMSE	Validation RMSE
Linear	4.87	4.96
Linear with $(T - 65)^2$	2.52	2.46

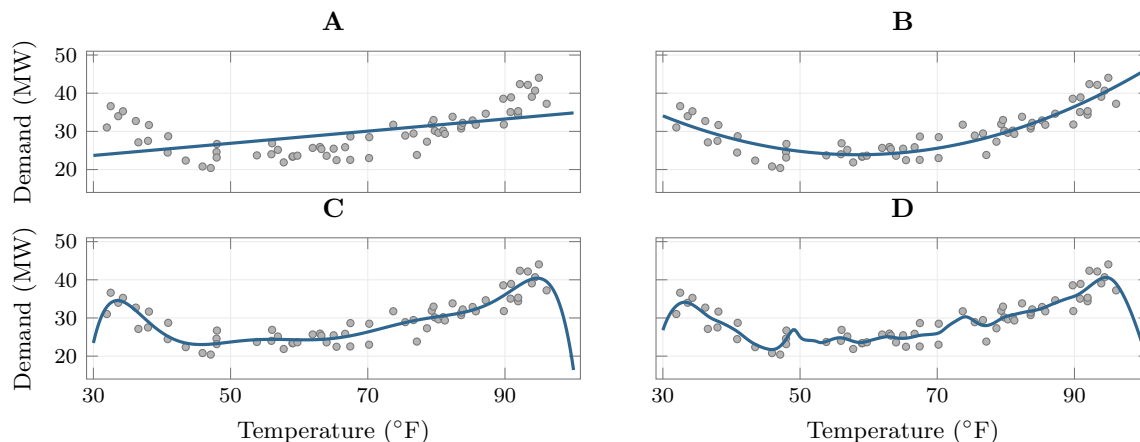
Which interpretation is best supported?

- A. The squared feature permits curvature. The displayed validation residuals show little systematic structure, and the lower validation RMSE supports the expanded model; it is nonlinear in T but linear in its coefficients.
- B. The expanded model is preferable because its training RMSE decreased; the validation results are not needed for this decision.
- C. Validation RMSE below training RMSE is by itself evidence that the validation outcomes entered model fitting.
- D. The curved fit is nonlinear in its coefficients, so it is no longer a linear regression model.

Question 3

2 points

The panels below use the same training data. Each blue curve is a global linear model, a quadratic model, a polynomial of degree 8, or a local linear regression with a small bandwidth.



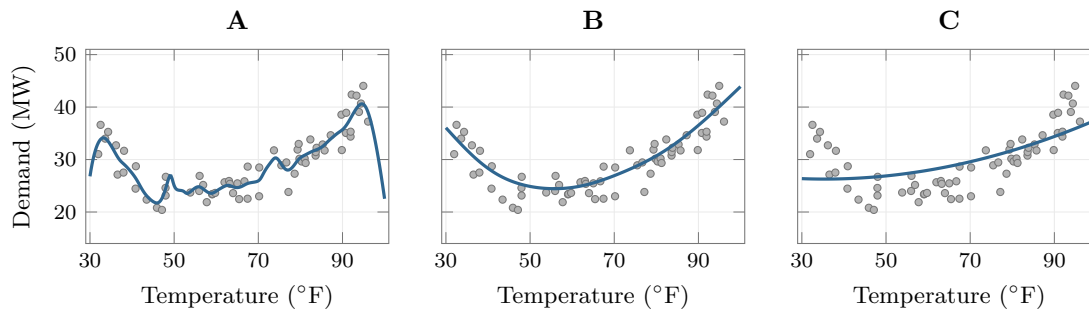
Which is the most plausible matching?

- A. A: quadratic; B: linear; C: polynomial of degree 8; D: local linear
- B. A: linear; B: local linear; C: polynomial of degree 8; D: quadratic
- C. A: linear; B: quadratic; C: polynomial of degree 8; D: local linear
- D. A: linear; B: quadratic; C: local linear; D: polynomial of degree 8

Question 4

2 points

Panels A–C show local linear regressions fit to the same data with three different bandwidths.



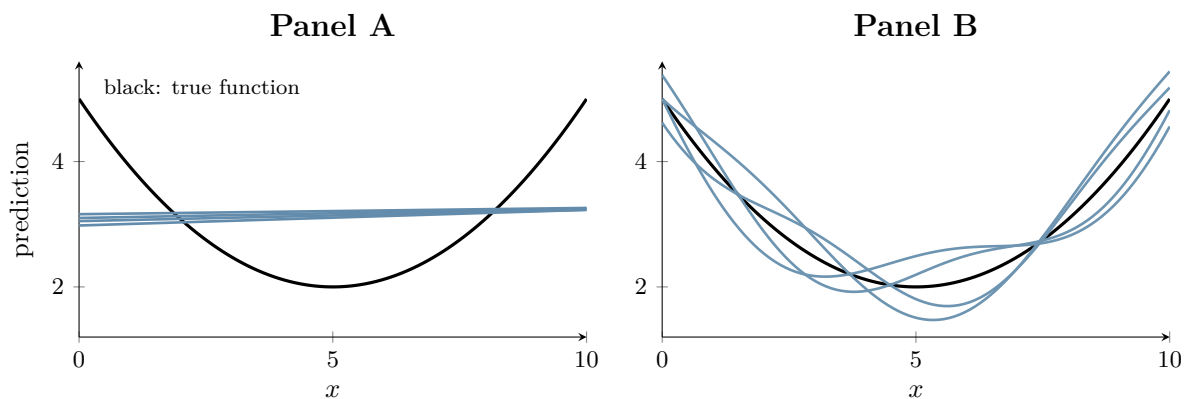
Which statement is correct?

- A. $h_A < h_B < h_C$. Panel A is most local and variable; panel C smooths most strongly and may have the most smoothing bias.
- B. Bandwidth changes only the estimated noise variance, not flexibility.
- C. $h_A > h_B > h_C$. Panel A is most variable because large bandwidths use fewer nearby observations.
- D. $h_A < h_B < h_C$. Panel C must have the smallest prediction error.

Question 5

2 points

The black curve in each panel is the true regression function. The blue curves are estimates obtained from several independently sampled training datasets using the same fitting procedure.



Which description is most appropriate?

- A. Panel A has low bias and high variance; panel B has high bias and low variance.
- B. Both procedures have low bias and low variance because every curve is smooth.
- C. The panels describe only irreducible error, not bias or variance.
- D. Panel A has high bias and low variance; panel B has lower bias and higher variance.

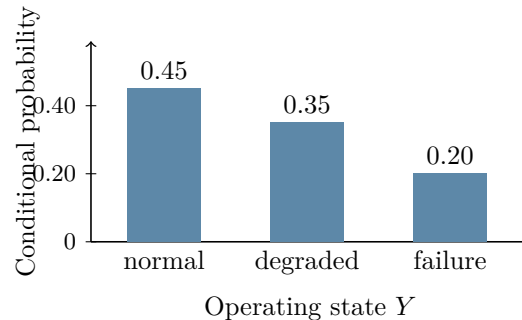
Setting 2: Robot condition classification

A campus monitors delivery robots using operating temperature and vibration. The questions below consider classification problems with three classes (multiclass classification) and two classes (binary classification), using class probabilities, decision boundaries, kNN, and support vector machines. The binary response is motor failure within seven days.

Question 6

2 points

For a robot with feature profile $X = x_0$, the true conditional probabilities of its operating state are shown below.



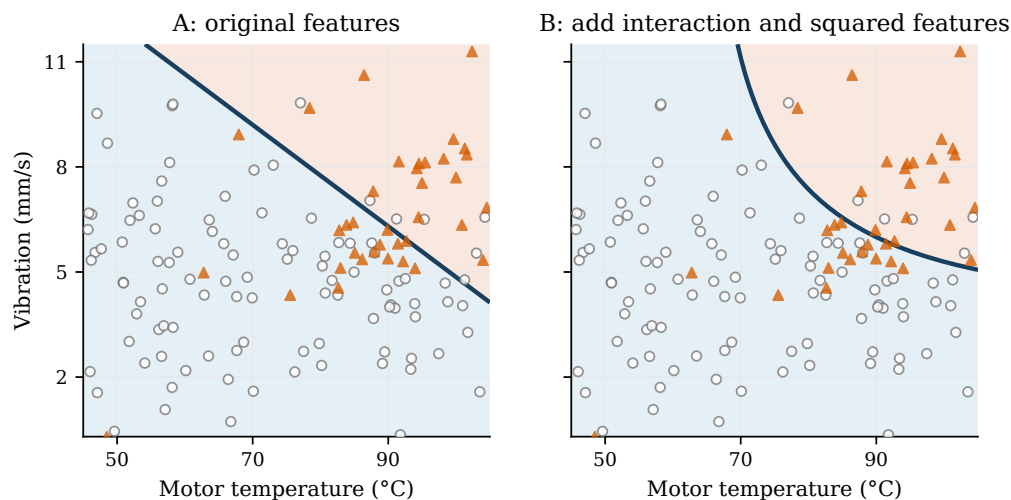
Under zero-one loss, what does the Bayes classifier predict, and what is its conditional probability of error at x_0 ?

- A. Degraded; error probability 0.65
- B. Do not classify, because no class exceeds probability 0.50
- C. Normal; error probability 0.55
- D. Normal; error probability 0.45

Question 7

2 points

Two logistic regression models are fitted to the robot data. Model A uses temperature and vibration. Model B additionally uses their interaction and squared terms. Their 0.50 decision boundaries are shown below. On the validation data, their log losses are 0.36 and 0.33, respectively.



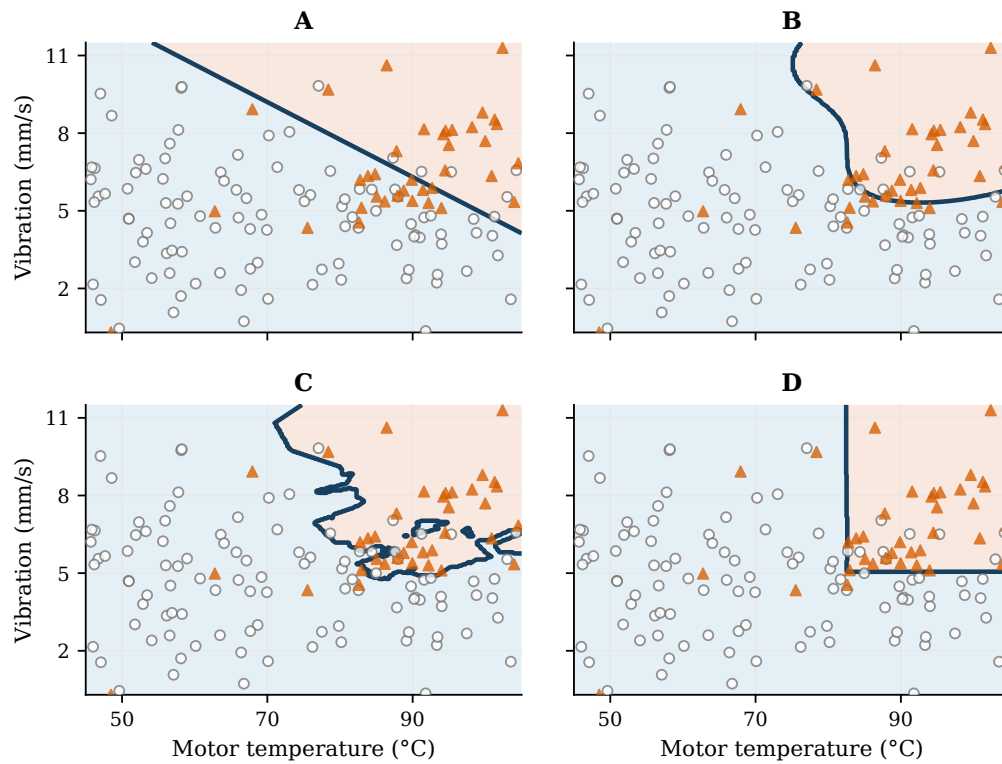
Which interpretation is most appropriate?

- A. The expanded features leave the represented boundary unchanged and only alter the classification threshold.
- B. The lower validation log loss proves that the interaction and squared terms are the scientifically correct mechanism.
- C. Model B remains linear in its coefficients, but the expanded features permit a nonlinear boundary; its lower validation log loss supports the expansion.
- D. A curved boundary implies that Model B must be an SVM with an RBF kernel rather than logistic regression.

Question 8

2 points

Each panel shows a classifier fitted to the same robot data. The methods are linear logistic regression, an SVM with an RBF kernel, kNN with $k = 5$, and a shallow decision tree.



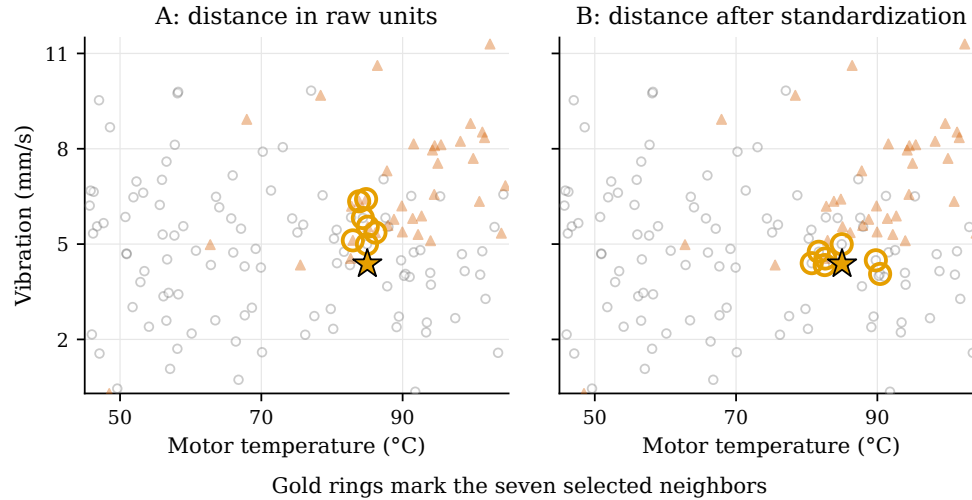
Which is the most plausible matching?

- A. A: logistic; B: RBF SVM; C: kNN; D: tree
- B. A: logistic; B: tree; C: kNN; D: RBF SVM
- C. A: tree; B: kNN; C: RBF SVM; D: logistic
- D. A: RBF SVM; B: logistic; C: tree; D: kNN

Question 9

2 points

The gold star marks the same query robot in both panels. Gold rings identify its seven nearest neighbors. Panel A uses the original units; panel B first standardizes both features.



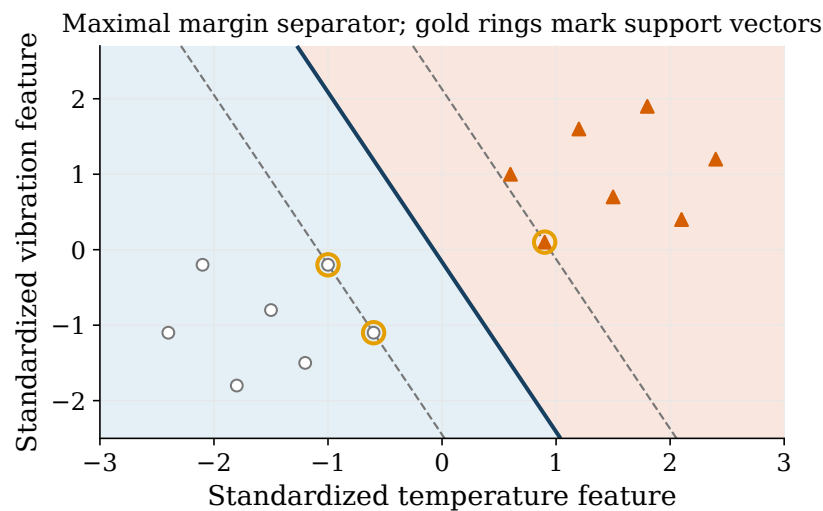
Why can the selected neighbors and prediction change?

- A. Euclidean distance depends on numerical scale, so measurement units determine how strongly each feature contributes to the geometry.
- B. Standardization guarantees neighbors with the query robot's class.
- C. Standardization changes the observed class labels.
- D. Standardization matters only for numerical optimization.

Question 10

2 points

For a linearly separable subset, the solid blue line is the hard margin SVM boundary, dashed lines mark the margins, and gold rings mark selected robots.



What is special about the observations marked with gold rings?

- A. They are errors that a hard margin SVM ignores.
- B. They are farthest from the boundary and determine its direction.
- C. They are support vectors. Moving one can change the fitted boundary, while small movements of observations away from the margin generally do not.
- D. They were randomly selected to estimate test error.

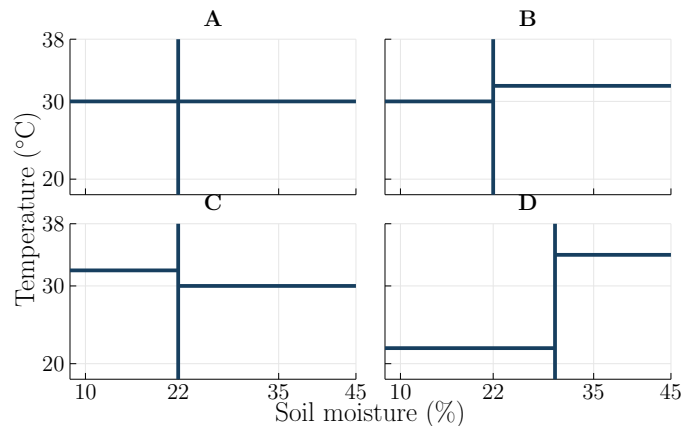
Setting 3: Garden water demand with trees and ensembles

A campus garden uses soil moisture and temperature to predict water demand on the next day. The questions below move from one regression tree to bagging, random forests, and boosting.

Question 11

2 points

Consider a tree that first splits on soil moisture below 22%. Within the drier region it splits on temperature below 30°C; within the moister region it splits on temperature below 32°C.



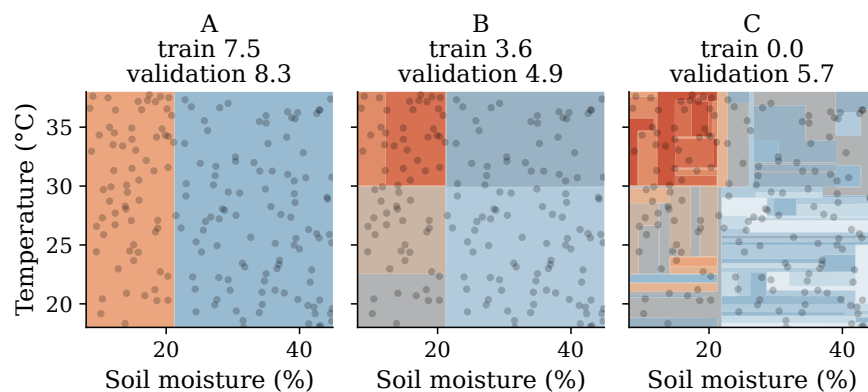
Which panel shows the induced partition of the predictor space?

- A. Panel A
- B. Panel B
- C. Panel C
- D. Panel D

Question 12

2 points

Trees A–C were fitted to the same garden data with progressively greater complexity. The titles report training and validation RMSE.



Which interpretation is best supported?

- A. Tree B has the lowest validation error; Tree C's perfect training fit and worse validation error provide evidence of overfitting.
- B. Pruning should use only training error, so Tree C is preferred.
- C. Tree A should be selected because the shallowest tree always wins.
- D. Tree C must generalize best because its training RMSE is zero.

Question 13

2 points

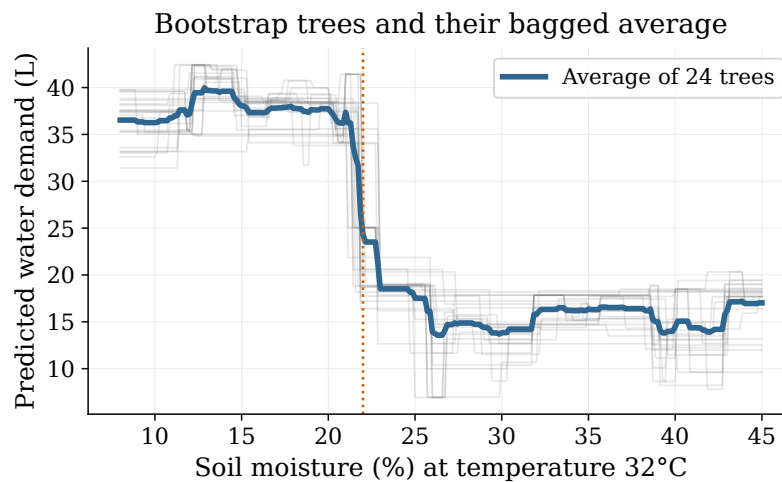
Two shallow regression trees fitted to different bootstrap samples of the same training data choose different early splits. Why can their partitions differ substantially?

- A. Bootstrap resampling changes the definition of water demand.
- B. The difference proves that one tree is not a regression model.
- C. Every bootstrap sample contains each training observation once.
- D. Trees can have high variance: modest sample changes can alter an early split and many later branches.

Question 14

2 points

Bagging fits 24 regression trees to bootstrap samples. The gray curves below are the individual predictions, and the blue curve is their pointwise average.



Why is the blue averaged prediction usually more stable than one gray tree?

- A. Averaging is equivalent to pruning each tree to a stump.
- B. Averaging predictions from trees with high variance can reduce variance, especially when their errors are not perfectly correlated.
- C. Bagging forces all trees to have the same splits.
- D. Averaging eliminates every source of bias.

Question 15

2 points

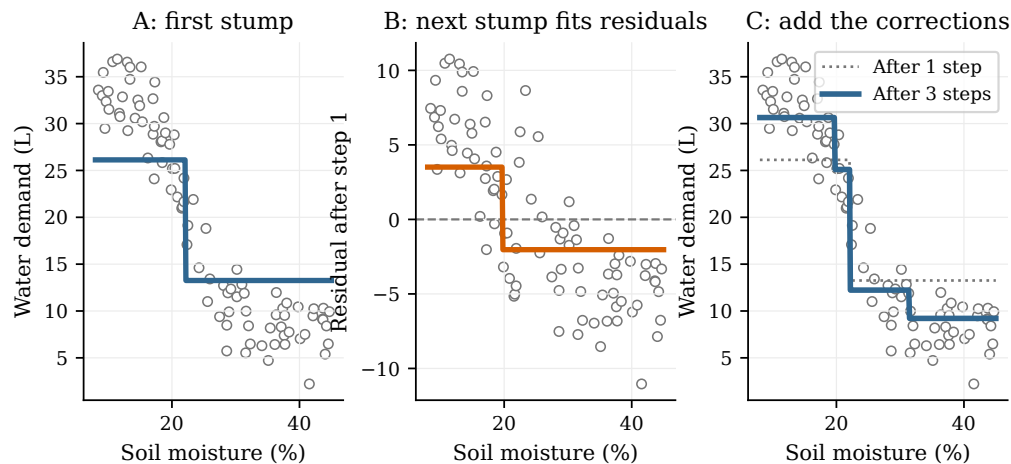
Recall that bagging considers all predictors as candidates at each split, whereas a random forest considers only a randomly selected subset as candidates. In one validation comparison, the mean correlation between tree predictions is 0.85 for bagging and 0.77 for the random forest. Why can the random forest benefit from the lower correlation?

- A. It uses more observations in every bootstrap sample.
- B. Feature subsampling can decorrelate trees, making averaging more effective even if each tree is not individually better.
- C. It removes all randomness from individual trees.
- D. Lower correlation guarantees that the ensemble is unbiased.

Question 16

2 points

The panels show early steps of boosting with squared error loss. Panel B fits a new stump to the residuals left by the first stump, and Panel C shows the accumulated prediction after adding corrections.



Which statement best distinguishes boosting from bagging?

- A. Boosting trains one tree and repeatedly changes its threshold.
- B. Both fit all trees independently to bootstrap samples and average.
- C. Bagging fits residuals sequentially; boosting fits the original response independently.
- D. Boosting builds trees sequentially to correct errors left by the current ensemble; bagging fits trees largely independently and averages.

Setting 4: Hospital readmission model development

A hospital develops a model to predict 30-day readmission at the moment of discharge. The questions below examine information available at prediction time, resampling with repeated patients, preprocessing, and final evaluation. The team uses cross-validation during development and reserves a final test set.

Question 17

2 points

The candidate predictors available by discharge are age, diagnoses recorded by discharge, laboratory values available before discharge, and length of stay. Which set can be used as stated for a prediction made at discharge?

- A. The predictors listed above, plus whether the patient is readmitted within 30 days
- B. The predictors listed above, plus medications prescribed at the patient's first appointment after discharge
- C. The predictors listed above, plus the number of emergency visits during the following 30 days
- D. The predictors listed above only

Question 18

2 points

The model will ultimately be used for patients not represented in the development data. To estimate performance for this goal during model development, how should the team construct the cross-validation folds?

- A. Put each patient's earliest visit in every validation fold.
- B. Keep every hospitalization from a patient in the same fold.
- C. Use the same observations for training and validation.
- D. Assign individual hospitalizations randomly, even when the same patient appears in multiple folds.

Question 19

2 points

The pipeline imputes missing values and standardizes predictors before fitting a classifier based on distance. How should these preprocessing steps be handled during cross-validation?

- A. Estimate them from the validation portion of each fold.
- B. Skip preprocessing whenever cross-validation is used.
- C. Estimate them within each training fold and apply them to that fold's validation observations.
- D. Estimate preprocessing quantities once from the complete dataset.

Question 20

2 points

The team evaluates 50 pipelines on the reserved test set, selects the one with the smallest test error, and reports that same error as final performance. What is the main problem?

- A. The procedure would be unbiased if the team compared only a small, fixed number of pipelines on the test set.
- B. There is no problem because the pipelines were fitted without using the test predictors; examining test errors does not affect selection.
- C. Taking the smallest test error makes the reported error conservative, because variable pipelines tend to look worse on test data.
- D. The test set has entered model selection, so the reported performance is likely optimistic.

Part II: Code and data analysis questions

Questions 21–40 use Settings 5–7, based on three real datasets.

Setting 5: Pottery clustering

The data contain 26 ancient pottery fragments from four sites, described by five chemical composition measurements (Al, Fe, Mg, Ca, and Na). The questions below use K-means and hierarchical clustering to examine feature choice, scaling, cluster selection, stability, and interpretation. **Site** is withheld from fitting and used only for interpretation.

```
features = ["Al", "Fe", "Mg", "Ca", "Na"]
X_raw = Pottery[features]
X_scaled = StandardScaler().fit_transform(X_raw)

results = []
for label, X in {"Raw": X_raw,
                 "Standardized": X_scaled}.items():
    for k in range(2, 7):
        fit = KMeans(n_clusters=k, n_init=50,
                     random_state=336).fit(X)
        results.append({
            "representation": label,
            "k": k,
            "WSS": fit.inertia_,
            "silhouette": silhouette_score(X, fit.labels_)
        })
```

Question 21

2 points

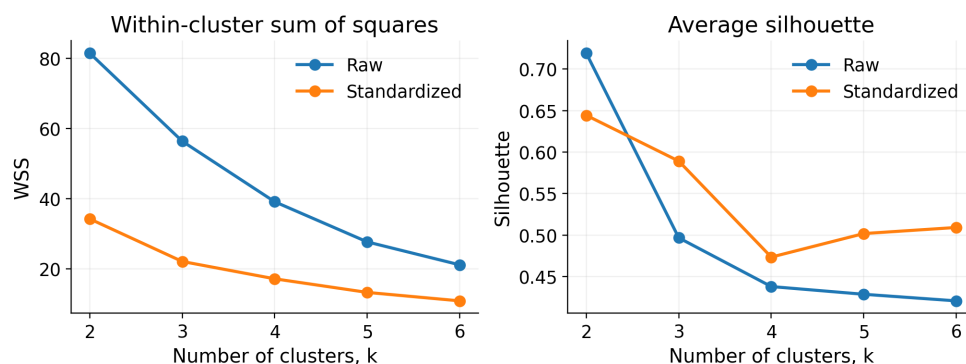
Why is **Site** not included in **features**?

- A. Including **Site** would force K-means to produce exactly four clusters.
- B. **Site** is contextual information rather than a chemical measurement; excluding it keeps the fit unsupervised and allows it to be used for interpretation afterward.
- C. **Site** must be the response variable because every machine learning problem requires a response.
- D. K-means cannot be run when a data frame contains any string column, even if that column is excluded before fitting.

Question 22

2 points

The code produces the following model selection curves.



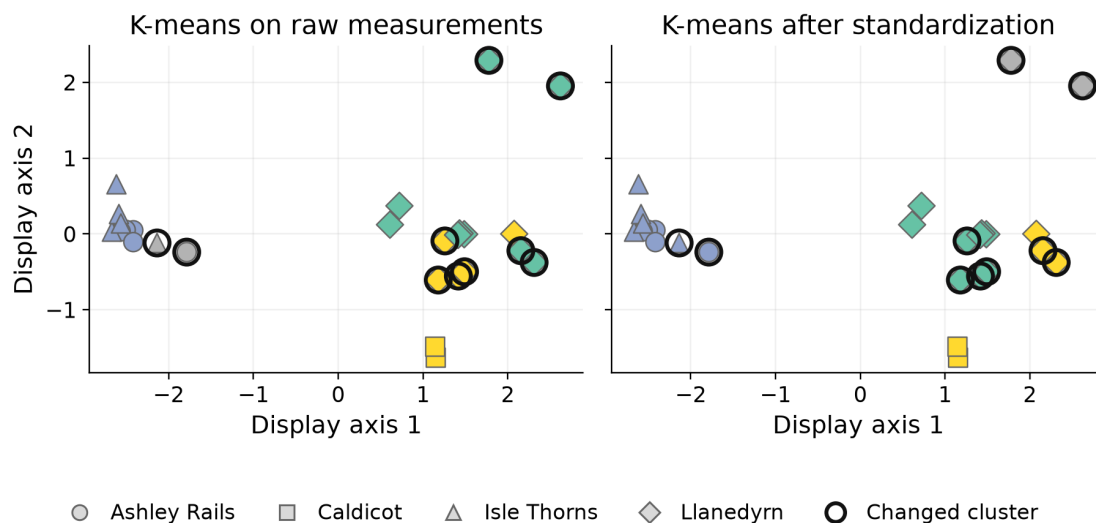
Based only on the silhouette curve for the standardized data above, which value of k is best supported?

- A. $k = 5$
- B. $k = 6$
- C. $k = 3$
- D. $k = 2$

Question 23

2 points

The next code fits $k = 4$ on the same five chemical measurements before and after standardization. Both panels use the same two-dimensional projection only for display. Marker shape indicates site, fill color indicates fitted cluster, and a bold black outline marks one of the ten fragments whose assignment changes. Matching colors represent aligned cluster labels.



What is the strongest conclusion?

- A. Feature scale is part of the clustering model because it changes distances and can change cluster assignments.
- B. The raw fit must be correct because the raw values are the original observations.
- C. The standardized fit must be correct because every feature has equal variance.
- D. Standardization only changes the plotting coordinates; it cannot change a fitted K-means partition.

Question 24

2 points

For the fitted standardized $k = 2$ solution, the following code compares the clusters with the unused Site column.

```
pd.crosstab(Pottery["Site"], cluster_labels)
```

Site	Cluster 1	Cluster 2
Ashley Rails	0	5
Caldicot	2	0
Isle Thorns	0	5
Llanedyrn	14	0

Which interpretation is best supported?

- A. This fitted solution broadly pairs Llanedyrn with Caldicot and Isle Thorns with Ashley Rails, rather than recovering four separate clusters, one for each site.
- B. Site caused the K-means algorithm to create these two clusters.
- C. Cluster 1 is intrinsically more important because it has the smaller numerical label.
- D. The clustering perfectly recovers all four excavation sites.

Question 25

2 points

The standardized data are fitted with K-means 30 times, using $k = 4$, one random initialization per fit, and a different seed each time. The results vary:

Across 30 runs	Minimum	Maximum
WSS	17.64	31.92
Silhouette	0.18	0.53

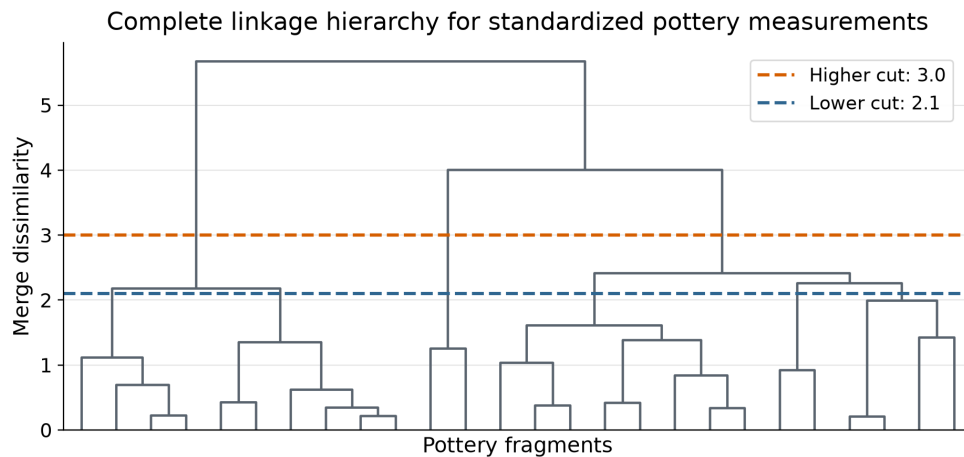
Which change most directly addresses the instability shown here?

- A. Increase k until all runs return the same WSS.
- B. Set `random_state=None` so the answer changes each time.
- C. Replace Euclidean distance with accuracy.
- D. Use multiple initializations and retain the run with the smallest K-means objective.

Question 26

2 points

Complete linkage hierarchical clustering is applied to the standardized chemical measurements. The same fitted dendrogram is shown with two possible cut heights.



Which interpretation is correct?

- A. Lowering the cut changes the linkage rule and therefore requires refitting the hierarchy from the original observations.
- B. The lower cut produces fewer clusters because more branches have already been merged below it.
- C. The higher cut produces three clusters and the lower cut produces six; the fitted hierarchy itself does not change.
- D. The number of clusters is fixed by complete linkage and cannot be changed by choosing a different cut height.

Setting 6: Auto regression

The Auto data contain 392 cars. The goal is to predict fuel economy, measured as miles per gallon (mpg), from recorded vehicle characteristics. The questions below compare baselines, linear and nonlinear regression, model selection, regularization, and resampling. A fixed 70/30 split into training and test sets is used: model choices are made within the training portion, and the test portion is reserved for final evaluation.

```
X = Auto[["horsepower"]]
y = Auto[["mpg"]]
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.30, random_state=336)

baseline = np.repeat(y_train.mean(), len(y_test))
linear = LinearRegression().fit(X_train, y_train)

print(rmse(y_test, baseline))
print(rmse(y_test, linear.predict(X_test)))
```

```
7.585
4.788
```

Question 27

2 points

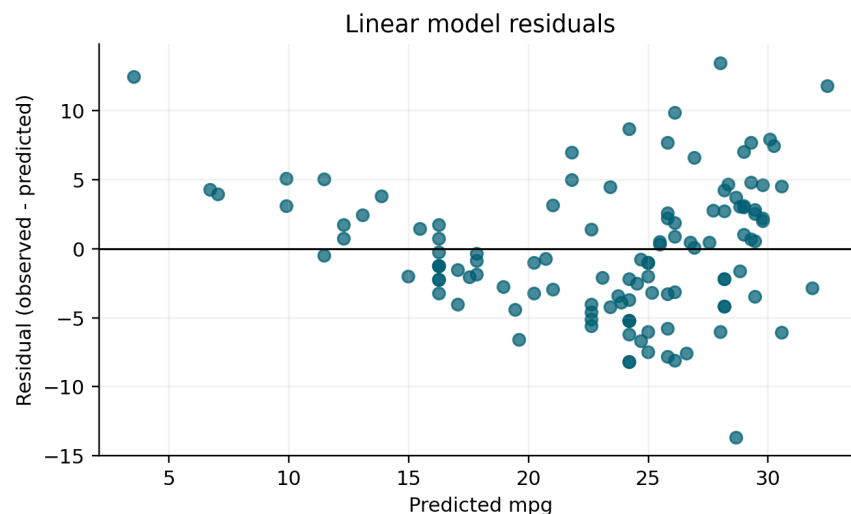
What is the most appropriate conclusion from the two test set RMSE values above?

- A. The value 4.788 is the percentage of variation in the test set explained by horsepower.
- B. On this test split, horsepower adds useful predictive information relative to the mean baseline, but the comparison does not establish that a linear model is adequate.
- C. The lower RMSE establishes that changing horsepower would causally change fuel economy by the fitted amount.
- D. Beating the baseline establishes that no nonlinear model could improve test prediction further.

Question 28

2 points

The residual plot below uses validation observations generated within the training portion of the data.



What does the systematic pattern most directly suggest?

- A. The pattern indicates changing residual spread, but the conditional mean has been represented adequately by a line.
- B. The deviations can be ignored because validation residuals need not average to zero locally in horsepower.
- C. The linear model is missing nonlinear structure in the conditional mean of mpg given horsepower.
- D. The residuals establish a causal effect of horsepower on fuel consumption, even without additional assumptions.

Question 29

2 points

The orange fitted curve below is nonlinear in horsepower. It is produced by supplying horsepower and horsepower squared to `LinearRegression`.



Why is this still called a linear regression model?

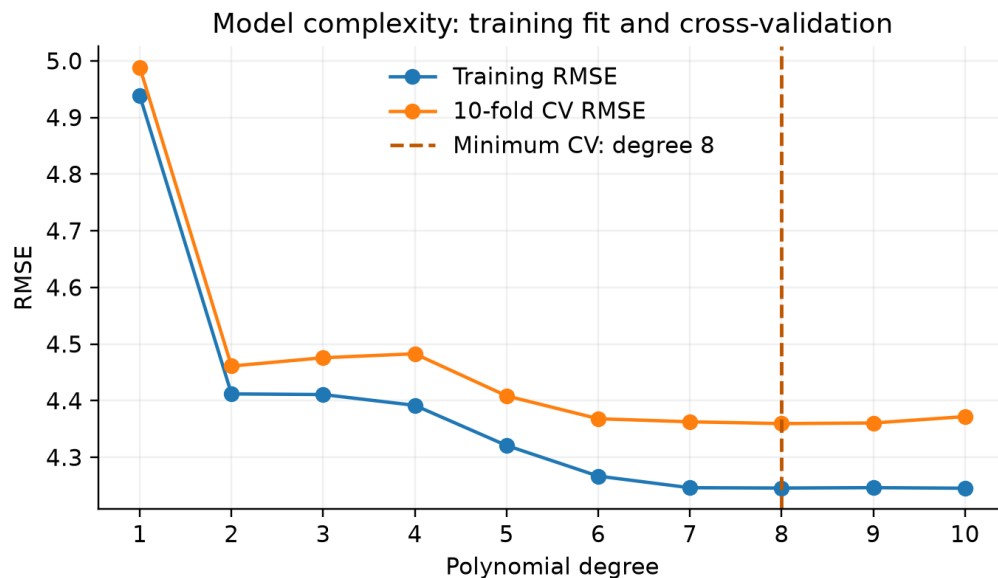
- A. Every regression model is linear after plotting.
- B. The model is linear in horsepower but nonlinear in its coefficients.
- C. The model is linear in its fitted coefficients, although it uses a nonlinear transformation of the original predictor.
- D. Squared features are ignored by `LinearRegression`.

Question 30

2 points

The following loop performs 10-fold cross-validation within the training set.

```
cv = KFold(n_splits=10, shuffle=True, random_state=336)
for degree in range(1, 11):
    model = make_pipeline(
        PolynomialFeatures(degree, include_bias=False),
        StandardScaler(),
        LinearRegression()
    )
    fold_mse = -cross_val_score(
        model, X_train, y_train, cv=cv,
        scoring="neg_mean_squared_error")
    cv_rmse[degree] = np.sqrt(fold_mse.mean())
```



Suppose the model selection rule was declared in advance: choose the degree with the smallest cross-validated RMSE, computed here as the square root of the mean validation MSE across folds. Which statement is correct?

- A. Degree 8 is selected using validation predictions generated within the training set; the test set has not made this decision.
- B. Degree 1 is selected because cross-validation cannot compare nonlinear feature expansions.
- C. Degree 10 is selected because training error must determine model complexity.
- D. Degree 2 is selected using the test observations.

Question 31

2 points

A second analyst proposes replacing cross-validation with the following loop.

```
for degree in range(1, 11):
    model = polynomial_pipeline(degree).fit(X_train, y_train)
    test_rmse[degree] = rmse(y_test, model.predict(X_test))
best_degree = min(test_rmse, key=test_rmse.get)
```

Why would this compromise the evaluation?

- A. The test set remains valid because the models themselves were fitted only on the training observations.
- B. Selecting on the test set is acceptable if the same test split is used consistently for every candidate degree.
- C. Reusing the test set for model selection leaks test information into the procedure and makes the eventual test assessment optimistic.
- D. The only problem is preprocessing: test predictors should be centered using their own mean before the models are compared.

Question 32

2 points

The seven numeric predictors are expanded to include all terms of degree 2 and pairwise interactions, giving 35 coefficients for the expanded features, excluding the intercept. Scaling and penalty selection are performed inside the training workflow.

Model	Test RMSE	Selected λ	Nonzero coefficients
OLS	2.907	—	35
Ridge CV	2.851	1.000	35
Lasso CV	2.799	0.0016	27

Which statement is supported by this output?

- A. The nonzero lasso coefficients identify causal effects of the corresponding vehicle characteristics.
- B. Both penalties slightly improve test RMSE here; ridge retains all coefficients, while lasso sets eight coefficients to zero.
- C. Ridge performs variable selection because all 35 coefficients are nonzero.
- D. Lasso has the largest test RMSE and therefore underfits.

Question 33

2 points

The test rows are sampled with replacement 3,000 times. In each bootstrap resample, both models are evaluated on the same sampled rows. The recorded difference is the linear model's RMSE minus the quadratic model's RMSE:

$$\text{RMSE}_{\text{linear}} - \text{RMSE}_{\text{quadratic}}.$$

```
np.quantile(bootstrap_differences, [0.025, 0.975])
```

```
array([0.182, 0.997])
```

Which interpretation is most appropriate?

- A. The interval proves the quadratic model is the true regression function.
- B. In this analysis based on resampling the test observations, the quadratic fit has lower RMSE because the interval is entirely positive.
- C. The interval shows that a degree 8 model must be used.
- D. The linear model has lower RMSE because 0 is outside the interval.

Setting 7: Homeownership classification

The cleaned 2019 American Housing Survey data contain 48,660 households. The goal is to predict homeownership, with OWNER=1 for an owner and 0 otherwise. The questions below examine class prevalence, confusion matrices, decision thresholds and costs, ROC AUC, calibration, model comparison, and subgroup evaluation. A stratified split into training and test sets preserves the owner rate; these are observational predictions, not estimates of causal effects.

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.30, random_state=336, stratify=y)

print(len(y_test))
print(y_test.mean())
print(max(y_test.mean(), 1 - y_test.mean()))
```

```
14598
0.631
0.631
```

Question 34

2 points

What does the last output represent?

- A. The test accuracy of always predicting the majority class
- B. The fraction of missing values
- C. The false positive rate on the test set
- D. The ROC AUC of logistic regression

Question 35

2 points

At threshold 0.50, the logistic regression confusion matrix is computed with rows = true class and columns = predicted class. Class 0 denotes no homeownership, and class 1 denotes homeownership.

```
confusion_matrix(y_test, logistic_probability >= 0.50)
```

	$\hat{Y} = 0$	$\hat{Y} = 1$
$Y = 0$	3093	2287
$Y = 1$	1344	7874

How many false negatives occur, and what is recall for owners?

- A. 2,287 false negatives; recall = $3093 / (3093 + 2287) = 0.575$
- B. 7,874 false negatives; recall = $(3093 + 7874) / 14598 = 0.751$
- C. 3,093 false negatives; recall = $(1344 + 7874) / 14598 = 0.631$
- D. 1,344 false negatives; recall = $7874 / (7874 + 1344) = 0.854$

Question 36

2 points

The same predicted probabilities are converted to classes using three thresholds.

Threshold	Accuracy	Owner recall	Specificity	Precision
0.30	0.724	0.954	0.329	0.709
0.50	0.751	0.854	0.575	0.775
0.70	0.712	0.660	0.800	0.850

If failing to identify a true homeowner is especially costly, which threshold is preferable among these choices, and what is the main tradeoff?

- A. The threshold is irrelevant because rankings never change.
- B. 0.30; it increases owner recall but lowers specificity.
- C. 0.70; it increases owner recall and specificity simultaneously.
- D. 0.50; it maximizes every reported metric.

Question 37

2 points

Suppose the application assigns cost 4 to each false negative and cost 1 to each false positive. The analyst applies that stated cost rule to the same three thresholds.

```
threshold_results["cost"] = (
    4 * threshold_results["FN"] + threshold_results["FP"])
threshold_results[["threshold", "FN", "FP", "cost"]]
```

Threshold	FN	FP	Cost
0.30	429	3608	5324
0.50	1344	2287	7663
0.70	3131	1075	13599

Which conclusion is appropriate?

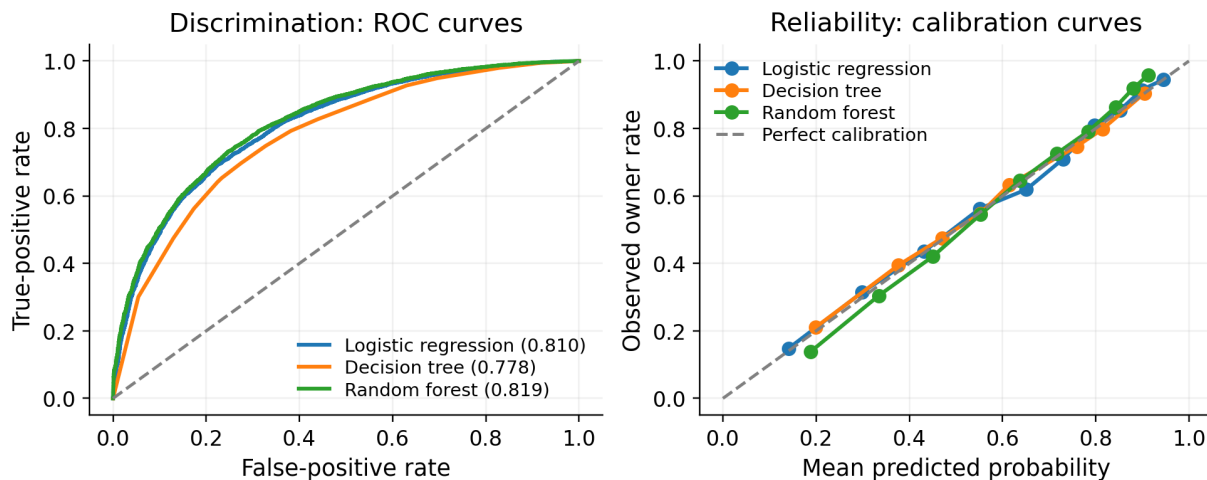
- A. Threshold 0.30 minimizes this particular stated cost among the three candidates; a different cost ratio could change the choice.
- B. Threshold 0.50 is always optimal because it is the default.
- C. The calculation is invalid because thresholds cannot be chosen using costs for the application.
- D. Threshold 0.70 is optimal because it has the fewest false positives.

Question 38

2 points

The following code evaluates ranking and calibration.

```
fpr, tpr, _ = roc_curve(y_test, probability)
observed, predicted = calibration_curve(
    y_test, probability, n_bins=10, strategy="quantile")
```



Which statement is correct?

- A. ROC AUC and calibration are identical because both use predicted probabilities.
- B. A model with the largest ROC AUC must also be perfectly calibrated.
- C. Calibration measures accuracy only at threshold 0.50.
- D. ROC AUC measures ranking across thresholds, while calibration compares predicted probabilities with observed frequencies.

Question 39

2 points

Three fitted models give the following test metrics.

Model	Accuracy	Balanced acc.	ROC AUC	Brier score
Logistic regression	0.751	0.715	0.810	0.168
Decision tree	0.729	0.694	0.778	0.180
Random forest	0.758	0.720	0.819	0.165

Which conclusion is best supported?

- A. The random forest should be chosen without further evaluation because having the best values for the metrics above establishes that it will be consistently superior.
- B. Logistic regression and the random forest are equivalent because all displayed differences are smaller than one percentage point.
- C. The random forest is best on all four reported metrics, but its advantage over logistic regression is modest and the table alone does not establish stability across resamples.
- D. The decision tree should be preferred because a single tree is easier to interpret, regardless of the stated predictive objective.

Question 40

2 points

The code below evaluates average probability predictions separately by the **BLACK** indicator.

```
comparison.groupby("BLACK").agg(
    observed_rate=("OWNER", "mean"),
    mean_logit=("logit_probability", "mean"),
    mean_forest=("forest_probability", "mean"))
```

Group	Observed	Logistic	Forest
Non-Black households	0.666	0.666	0.661
Black households	0.421	0.405	0.438

Using the displayed rounded values, which conclusion follows from these subgroup averages?

- A. The forest should be used because overall ROC AUC should override any discrepancy in subgroup probability averages.
- B. Logistic regression more closely reproduces the two displayed group rates in this particular summary, despite the forest's slightly stronger overall metrics.
- C. The models are indistinguishable within each group because their overall accuracy values are similar.
- D. Matching these two group averages would be sufficient to establish equalized odds and every other common fairness criterion.