Practice Final Exam

MSE 125: Applied Statistics — Spring 2026

About this practice exam

This practice exam mirrors the structure of the real final: **3 sections, 100 points, 90 minutes**. The data, scenarios, and specific numbers differ from the real exam, but every item exercises the same skill as its real-exam counterpart.

Take this under simulated exam conditions: 90 minutes timed, no devices, no AI, no notes other than the formula strip provided in Section 1. Then check your answers against the solutions version.

- **Section 1** (25 points, ∼22 min): tool literacy. 8 multiple-choice + 5 short fill-in. A formula strip is provided at the head of the section.
- **Section 2** (35 points, ∼33 min): interpretation and EDA. Three problems with figures.
- **Section 3** (40 points, ∼35 min): diagnose and supervise. Three longer problems, starting with the AI code review.

The 8 unit practice quizzes remain your primary study resource. This practice final adds cumulative-cross-lecture practice and exposure to the two novel archetypes (AI code review, diagnose-the-phenomenon) that appear only on the final.

No aids permitted other than the formula strip in Section 1. Just your brain and your pencil.

Begin when ready. Set a 90-minute timer.

Name: ____________________ **SUNet ID:** ____________________

Section 1 — Tool literacy (25 pts, ∼22 min)

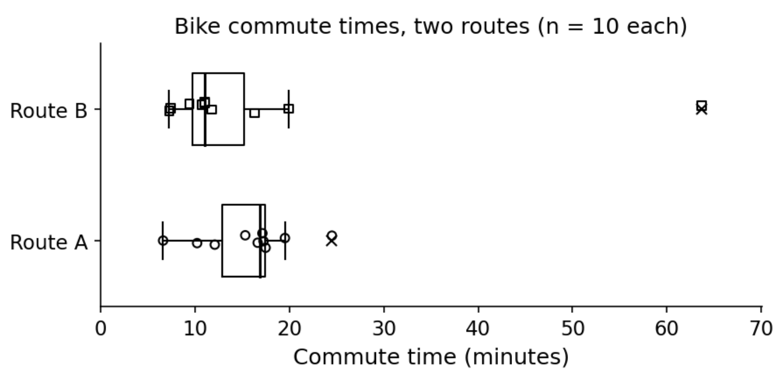
Formula strip (Section 1 only):

Bonferroni cutoff $= \alpha/m$ $\mathbb{E}$[false positives, all-null] $= m\alpha$ Recall $= \frac{TP}{TP+FN}$

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

$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$ Residual $= y - \hat{y}$ $|t| = \frac{|\hat{\beta}|}{\text{SE}(\hat{\beta})}$ $z_{0.975} \approx 1.96$

1. (2 points) **Test selection.** You measure bike commute times (minutes) on two routes. The boxplots below show $n = 10$ observations per route; both distributions are right-skewed and Route B contains one clear outlier.



   Which test is most appropriate for comparing the two routes' mean commute times?

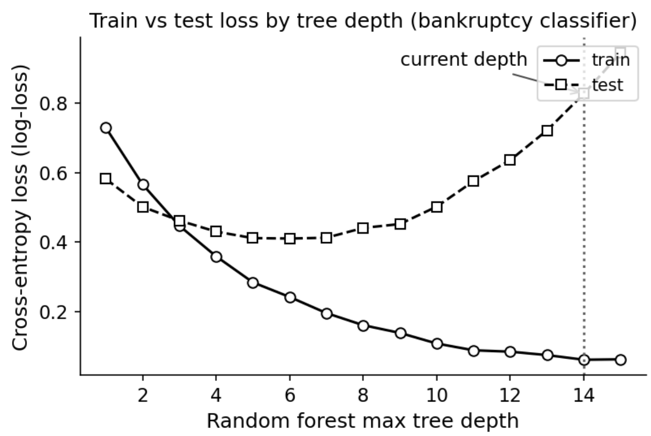
   A. Two-sample t-test

   B. Permutation test on the difference in means

   C. Sign test on each route's distribution

   D. Bootstrap CI for each route's mean, separately

2. (2 points) **Identify under/overfit.** The plot shows training and test cross-entropy loss for a random forest as a function of `max_depth`. The model is currently trained at the marked depth.



The model is currently:

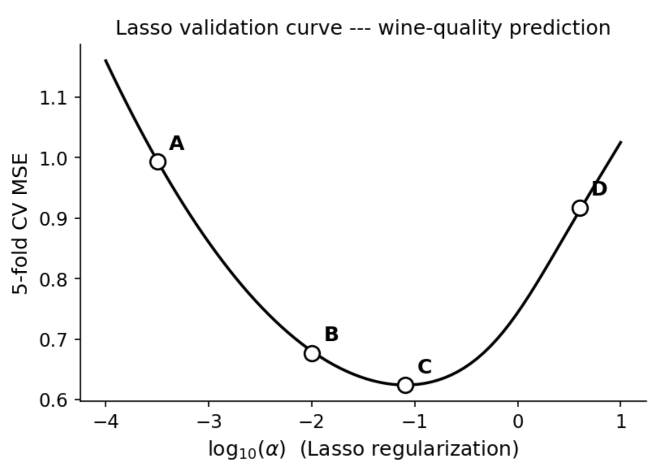
A. Underfitting (high training and test loss)

B. Good fit (test loss at its minimum)

C. Overfitting (training loss very low, test loss rising)

D. Cannot tell from this plot

3. (2 points) **Best point on a validation curve.** The plot shows mean 5-fold CV mean squared error for Lasso as a function of $\log_{10}(\alpha)$ on a wine-quality prediction task. Four candidate operating points A, B, C, D are marked.



Which point should you choose for deployment?

A. A (smallest α)

B. B (small α)

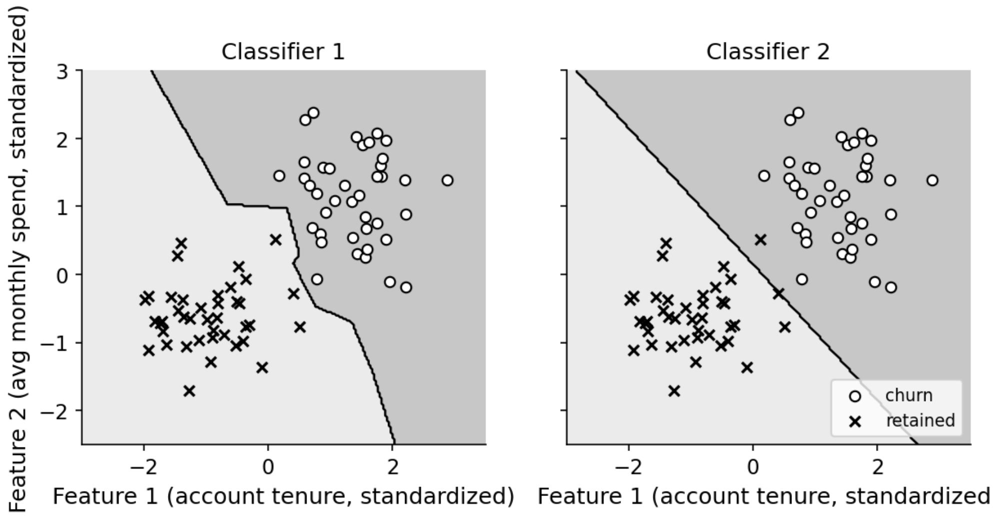
C. C (CV-error minimum)

D. D (largest α)

4. (2 points) **Cross-validation protocol.** You have a panel of standardized test-score records for 200 schools across 6 academic years. You want to estimate how well your model will generalize to a *new school* the district has not yet onboarded. Which CV protocol gives an honest estimate?

A. Random 5-fold cross-validation across all (school, year) records
B. Walk-forward validation by year
C. Grouped CV by school (hold out all records from a school in each fold)
D. Stratified k-fold, stratified by year

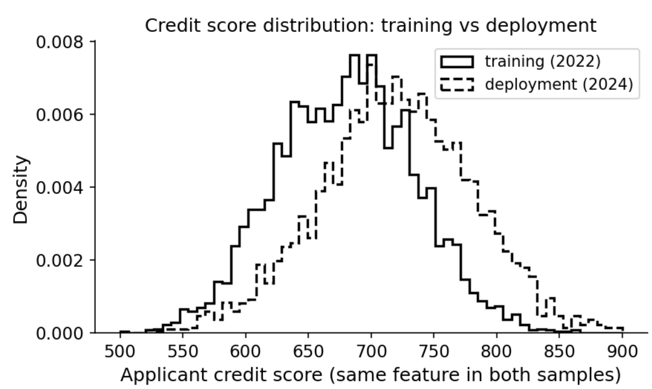
5. (2 points) **Decision boundary identification.** The two panels below show classifier decision boundaries on the same 2D customer-churn scatter. Classifier 1 produces a jagged, locally-shaped boundary; Classifier 2 produces a single straight line.



Which panel was produced by which model?

A. Classifier 1: linear SVM; Classifier 2: k-nearest neighbours
B. Classifier 1: k-nearest neighbours; Classifier 2: linear SVM
C. Both: k-nearest neighbours, with different k values
D. Both: linear SVMs, with different features

6. (2 points) **Identify distribution shift.** The plot below shows kernel-density estimates of the same input feature (applicant `credit_score`) at two times: the training data (collected 2022) and the deployment data (collected 2024, after a policy change in upstream credit reporting). The two curves have visibly different shapes.



What does this plot evidence?

A. Covariate shift ($P(X)$ has changed)

B. Label shift ($P(Y)$ has changed)

C. No shift; differences are within sampling noise

D. Cannot tell — need to see $P(Y \mid X)$

7. (2 points) **Lasso vs ridge vs OLS.** A colleague says: "I have 40 candidate features. I am required to keep all of them in the model (a contract obligation). The OLS fit is unstable across data refreshes — coefficients flip sign from month to month. I need a model with stable, lower-variance coefficient estimates that still uses every feature." Which model best supports this goal?

   A. Ordinary least squares (OLS) with all 40 features

   B. Ridge regression

   C. Lasso regression

   D. Decision tree with all 40 features available

8. (2 points) **Plot type for the question.** You want to know whether the relationship between daily temperature (continuous) and total bike-share rentals (continuous) is different on weekdays vs weekends. Which plot type is best?

   A. Histogram of rentals, one panel per weekday/weekend

   B. Bar chart of mean rentals by temperature bin

   C. Scatter plot of rentals against temperature, with weekday/weekend distinguished by marker or panel

   D. Box plot of rentals, one box per weekday/weekend

9. (2 points) **Confusion matrix → precision.** A radiology AI classifier flags scans as "suspicious for cancer." On 1,000 screening scans (100 actual cancer cases, 900 cancer-free), the model produces:

	Predicted cancer	**Predicted not cancer**
Actual cancer	27	73
Actual not cancer	3	897

   Compute the model's precision (round to 2 decimal places):

   Precision = ____________

10. (2 points) **Bonferroni cutoff.** You run $m = 50$ independent hypothesis tests and want to control the family-wise error rate at $\alpha = 0.05$. The Bonferroni per-test cutoff is:

    per-test α = ____________

11. (1 point) **Expected false positives.** You run $m = 400$ hypothesis tests at $\alpha = 0.01$, and *every null hypothesis is true* (no real effects). How many false positives do you expect, on average?

    $\mathbb{E}[\mathrm{FP}]$ = ____________

12. (2 points) **Bootstrap CI → reject or fail to reject.** An A/B test of a new checkout flow estimates the difference in conversion rate (new − old). A bootstrap of $B = 10{,}000$ resamples gives a 95% confidence interval of $[-0.4\%, +1.2\%]$.

    At $\alpha = 0.05$, do you reject the null hypothesis H_0: difference = 0? Circle one and give a one-sentence reason.

    Reject **Fail to reject**

13. (2 points) **Regression coefficient distinguishable from zero?** A regression of a continuous outcome on a single predictor yields:

Predictor	**Coefficient**	**Std error**
`predictor_x`	1.2	0.9

    Is this coefficient statistically distinguishable from zero at $\alpha = 0.05$? Circle one and give a one-sentence reason.

    Yes **No**

Section 2 — Interpretation & EDA (35 pts, ∼33 min)

14. (12 points) **Regression coefficient table (bike-share demand).** A bike-share operator fits a multivariable linear regression of daily ridership (rides per day) on weather and calendar features from one year of operations. The table reports estimates, standard errors, and p-values; the reference category for season is `Winter`. A side table gives the standard deviations of two continuous predictors in the fitting sample.

Predictor	**Estimate**	**Std err.**	**p-value**
Intercept	1,840	110	< 0.001
`temperature` (per °F)	62.4	3.8	< 0.001
`season_Spring`	320	85	< 0.001
`season_Summer`	540	96	< 0.001
`season_Fall`	210	88	0.018
`humidity` (per %)	−7.8	1.6	< 0.001
`is_holiday`	−380	140	0.007

Predictor	**SD**
`temperature`	17 °F
`humidity`	14 %

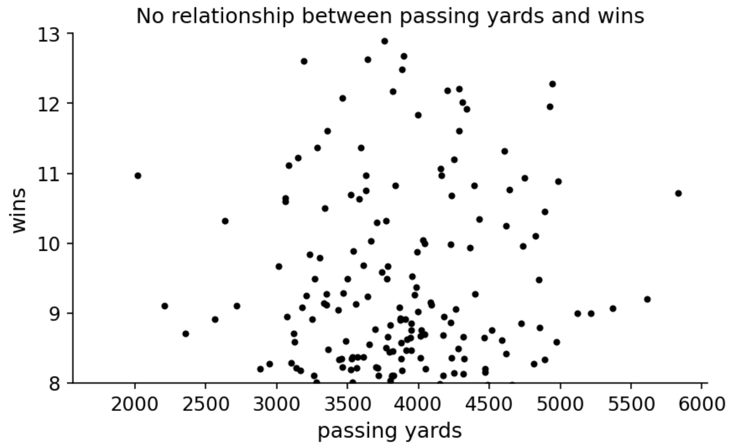
(a) (3 points) Interpret the coefficient on `temperature` in one sentence, with units.

(b) (3 points) Interpret the coefficient on `season_Summer` relative to the omitted reference category, in one sentence.

(c) (3 points) Of `temperature` vs `humidity`, which predictor moves predicted ridership more per one-standard-deviation change in the predictor? Show your work.

(d) (3 points) A colleague proposes adding `previous_day_ridership` as a predictor. Will this help predict ridership **for tomorrow, in a forecast made the night before**? Why or why not?

15. (12 points) **EDA plot critique.** A junior analyst produces the scatter plot below from a team-season dataset (one point per team-season across many years of NFL play) and writes the caption: "*There is no relationship between passing yards and wins.*"



(a) (6 points) Identify **three specific problems** with the plot itself (not the conclusion). One short sentence per problem.

Problem 1:

Problem 2:

Problem 3:

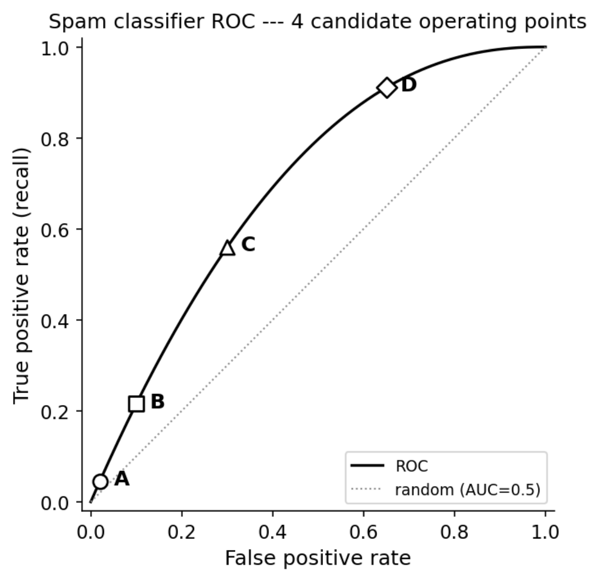
(b) (3 points) What plot type, transformation, or rendering choice would better answer the underlying question? One-sentence justification.

(c) (3 points) Do you trust the analyst's conclusion that there is "no relationship" between passing yards and wins? Why or why not?

16. (11 points) **Spam classifier.** An email provider fits a classifier on 2022 messages to predict whether each new message is spam. At the default threshold $\hat{p} = 0.5$, the confusion matrix on a held-out 2022 test set of 1,000 messages (300 actual spam, 700 legitimate) is:

	Predicted spam	**Predicted not spam**
Actual spam	170	130
Actual not spam	210	490

The ROC curve below shows the model's full TPR-vs-FPR trade-off. Four candidate operating points A, B, C, D are labeled.



(a) (3 points) Which of A, B, C, D corresponds to the displayed confusion matrix? Circle one.

A B C D

(b) (4 points) The email provider's product team says "customers complain loudly when legitimate emails land in their spam folder — we need to **minimize false positives**." Should they move the threshold **up** or **down**? What does that trade against?

(c) (4 points) The model was trained on 2022 spam patterns. By 2025, spammers have changed their tactics in response to public spam-detection write-ups. **Name the phenomenon** and recommend **one concrete action** the provider should take before continuing to rely on this model.

Section 3 — Diagnose & supervise (40 pts, ∼35 min)

17. (15 points) **AI code review.** A bike-share operator asks an AI agent: "*Predict, for each day, whether ridership will exceed station capacity.*" The dataset (`bike-ridership.csv`) has one row per day, sorted by date, with weather and calendar features and a binary target `exceeds_capacity` (the operator says exceedance is uncommon, somewhere around 5% of days). The agent returns the following code:

```
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import KFold, cross_val_score

df = pd.read_csv("bike-ridership.csv").sort_values("date")

features = ["temp_f", "humidity", "is_weekend", "is_holiday"]
X = df[features]
y = df["exceeds_capacity"]

cv = KFold(n_splits=5, shuffle=True, random_state=0)
model = RandomForestClassifier(n_estimators=100, random_state=0)

scores = cross_val_score(model, X, y, cv=cv, scoring="accuracy")
print(f"CV accuracy: {scores.mean():.3f}")
print("Model accuracy is above 0.90 -- looks good, ship it.")
```

(a) (6 points) Identify the **two bugs** in this code. Name each (a short label is fine) and quote the offending line.

Bug 1 name:

Offending line:

Bug 2 name:

Offending line:

(b) (4 points) On a future week with similar weather to the training data, what fraction of the days when ridership actually exceeds capacity does the deployed model correctly flag? Circle one and give a one-sentence reason.

High (≈ 1) **Moderate (≈ 0.4–0.7)** **Near zero**

(c) (3 points) For each bug, write the **one-line fix** (pseudocode is fine).

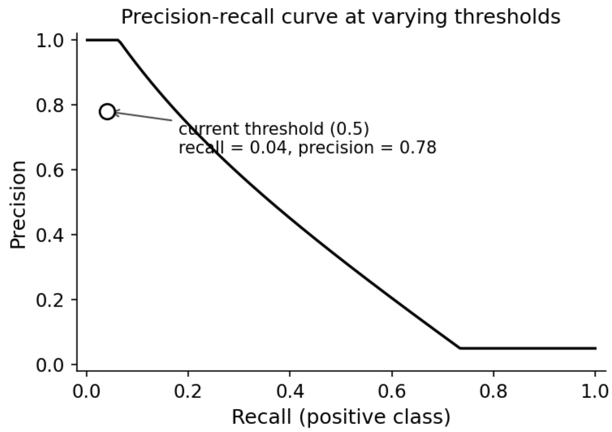
Fix 1:

Fix 2:

(d) (2 points) Name **one sanity check** that would catch *either* bug before deployment.

18. (12 points) **Diagnose the phenomenon.** For each scenario, name **two plausible causes** of the observed phenomenon. For each cause, write one sentence on **how you'd check it**.

   (a) (6 points) *"I fit a logistic regression classifier. Its accuracy on a held-out set is 0.96, but its recall on the positive class is 0.04."* The precision-recall curve at varying classification thresholds is shown below; the operating point at the default threshold $\hat{p} = 0.5$ is marked.



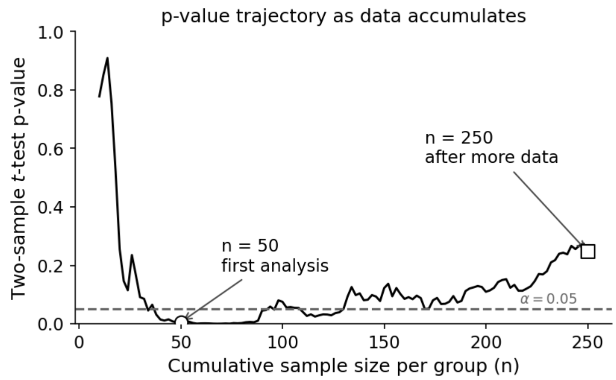
Cause 1:

Check:

Cause 2:

Check:

(b) (6 points) *"We ran a two-sample t-test on a difference in means and got* $p = 0.03$ *on* $n = 50$ *observations per group. We collected 200 more observations per group; the p-value is now around 0.25."* The trajectory of the p-value as the sample grows is shown below.



Cause 1:

Check:

Cause 2:

Check:

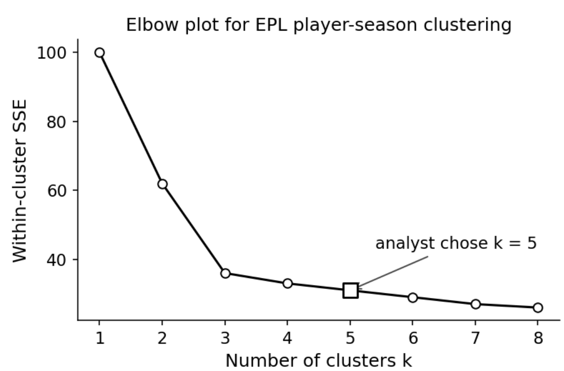
19. (12 points) **Unsupervised interpretation & decision (EPL player clustering).** An analyst clusters English Premier League player-season records with k-means at $k = 4$ on five standardized per-90-minute features: goals (G), assists (A), tackles (TKL), passes (PASS), and saves (SV). The cluster centroids (in standardized units, where 0 is league average and +1 is one SD above) and cluster sizes are below.

Cluster	**G**	**A**	**TKL**	**PASS**	**SV**	**Size**
1	+1.8	+0.3	−0.6	+0.2	−0.2	35
2	+0.1	+1.4	−0.2	+1.3	−0.2	80
3	−0.6	−0.5	+1.3	+0.5	−0.2	90
4	−0.4	−0.4	−0.4	−0.6	+2.5	18

(a) (4 points) Give each cluster a one-phrase plain-English name based on its centroid. A few words each is fine.

Cluster 1: ______________
Cluster 2: ______________
Cluster 3: ______________
Cluster 4: ______________

(b) (4 points) The analyst chose $k = 5$. The elbow plot below shows within-cluster sum of squares (WSS) as a function of k. Does the plot justify $k = 5$? If not, what k would you choose, and why?



(c) (4 points) Two runs of k-means at different random seeds, side by side, are shown below: many players land in different clusters across the two runs. The team's recruiting department wants to publish an internal shortlist of players in "Cluster 2" as midfielder targets. What does this comparison reveal, and what should the analyst do before publishing?