UT-ECE Data Science – Extended Final
Comprehensive Solution Manual (TA Edition)

Course Staff

Spring 2025

Scoring Philosophy

Use evidence-based grading: correctness, rigor, reproducibility, and interpretation quality. Prefer transparent assumptions and explicit limitations over overconfident claims.

Q1. Lifecycle and Problem Framing

High-quality answer includes:

- clear business target (e.g., early identification of migration propensity),
- operational metric (AUC, recall@k, calibration, fairness constraints),
- lifecycle phases: framing -> collection -> validation -> modeling -> deployment -> monitoring,
- risk register (leakage, drift, policy shift, proxy bias).

Q2. Python/EDA

Expected components:

- dtype audit, null profile, duplicate checks, range sanity checks,
- at least six meaningful visualizations with non-trivial interpretation,
- modular preprocessing function with unit tests.

Q3. Scientific Studies and Inference

Key grading points:

- distinguishes observational limits from causal claims,
- states assumptions for CI/hypothesis testing,
- interprets p-values and confidence intervals correctly.

Example framing:

- Null: $H_0 : \Delta\mu = 0$ for migration propensity proxy between cohorts.
- Use two-sample test with variance assumptions checked.

Q4. Visualization and Storytelling

Strong solution:

- KPI definitions tied to stakeholder decisions,
- perceptual design rationale (position/length over area/color where possible),
- explicit warning about misleading axis truncation or inappropriate color scales.

Q5. SQL Advanced Querying

Canonical moving-average query pattern:

```
WITH citation_velocity AS (
  SELECT UserID, Country_Origin, Year, Research_Citations,
         AVG(Research_Citations) OVER (
           PARTITION BY Country_Origin
           ORDER BY Year
           ROWS BETWEEN 2 PRECEDING AND CURRENT ROW
         ) AS moving_avg_citations
  FROM Professionals_Data
)
SELECT *, DENSE_RANK() OVER (
  PARTITION BY Country_Origin ORDER BY moving_avg_citations DESC
) AS country_rank
FROM citation_velocity;
```

Additional SQL expectations:

- percentile bucketing (e.g., NTILE or percentile window),
- cohort/transition query via CTE.

Q6. Leakage and Big-Data Architecture

Leakage decisions:

- `Visa_Approval_Date`: direct leakage (post-outcome),
- `Last_Login_Region`: potential temporal leakage,
- `Passport_Renewal_Status`: possible temporal proxy leakage,
- `Years_Since_Degree`: acceptable if computed pre-inference.

Architecture answer (acceptable): Bronze/Silver/Gold tables, feature store with point-in-time joins, online/offline feature parity, periodic drift checks.

Q7. Regression and Elastic Net

For

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda_1 \sum_j |\theta_j| + \frac{\lambda_2}{2}\sum_j \theta_j^2,$$

$$\nabla_{\theta_j} J = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} + \lambda_1 \partial|\theta_j| + \lambda_2\theta_j,$$

$$\partial|\theta_j| = \begin{cases} +1 & \theta_j > 0 \\ -1 & \theta_j < 0 \\ [-1,1] & \theta_j = 0 \end{cases}$$

Q8. Optimization

Ravine intuition: steep curvature in one axis and shallow curvature in another causes SGD oscillation.

Momentum:

$$v_t = \beta v_{t-1} + \eta g_t, \quad \theta_{t+1} = \theta_t - v_t$$

Damps sign-flipping gradients and accelerates consistent directions.

Adam: first and second moments with bias correction, parameter-wise scaling.

Q9. Model Family Comparison

Minimum expected protocol:

- fixed train/validation/test split with stratification,
- CV and hyperparameter tuning for each model family,
- metric table with at least AUC, F1, calibration/error analysis,
- interpretability discussion.

Q10. Dimensionality Reduction

PCA explained variance ratio:

$$\mathrm{EVR}_k = \frac{\lambda_k}{\sum_i \lambda_i}$$

where λ_k is variance captured by component k.

Q11. Clustering

K-Means elbow rationale: WCSS decreases monotonically with K, but marginal gain diminishes.

Density alternative: DBSCAN robustness to non-spherical clusters and noise points.

Q12. Neural Networks and Sequence Models

Expected answer characteristics:

- clear architecture choice and training setup,
- baseline comparison against classical model,
- overfitting control (dropout/early stopping/regularization).

Q13. LMs and LLM Agents

Strong answer includes:

- agent workflow (plan -> retrieve -> reason -> verify),
- evaluation: faithfulness, hallucination, safety,
- governance boundaries and fallback logic.

Q14. Ethics and Fairness

Expected:

- subgroup metrics (e.g., by country/education),
- recognition of historical policy bias and proxy discrimination,
- mitigation and human override policy.

Q15. Calibration and Threshold Policy

Expected answer components:

- reliability plot (calibration curve) with interpretation,
- at least one probabilistic calibration metric (Brier score and/or ECE),
- threshold policy from two objectives:
  - maximize F1,
  - minimize asymmetric expected cost.

Grading note: threshold choice must be justified by task costs, not by arbitrary default 0.5.

Q16. Drift Detection and Monitoring

Expected answer components:

- two-window split design (preferably temporal),
- numeric feature drift ranking via PSI,
- one categorical drift signal (e.g., JS divergence),
- clear trigger policy for warning/critical events.

Reference interpretation of PSI:

- PSI < 0.10: low drift,
- 0.10–0.25: moderate drift,
- PSI >= 0.25: high drift requiring intervention.

Q17. Counterfactual Recourse

Expected answer components:

- actionable feature set with practical constraints,
- minimal-change search per candidate near decision boundary,
- recourse success rate and per-feature effort summary,
- discussion of realistic and ethical intervention boundaries.

Grading note: penalize unrealistic interventions (e.g., impossible immediate changes) if not explicitly acknowledged.

Q18. Temporal Backtesting and Rolling Validation

Expected answer components:

- Explicit chronological split strategy with rolling folds.
- If no valid time field exists, a documented fallback ordering strategy.
- Fold-wise metrics (at minimum AUC and F1), with decay measured relative to the first fold.
- Drift-aware interpretation (e.g., mean PSI per fold or equivalent drift proxy).

Minimum acceptable artifacts:

- `q18_temporal_backtest.csv`
- `q18_temporal_degradation.png`

Grading note: if fallback chronology is used, students must explicitly justify why and state threat-to-validity impact.

Q19. Uncertainty Quantification

Expected answer components:

- Split-conformal or equivalent calibrated uncertainty procedure.
- Empirical coverage at multiple confidence levels.
- Interval width analysis and under-coverage reporting.
- Practical handling policy for low-confidence predictions.

Minimum acceptable artifacts:

- `q19_uncertainty_coverage.csv`
- `q19_coverage_vs_alpha.png`

Grading note: students lose points if they report confidence levels without empirical coverage validation.

Q20. Fairness Mitigation Experiment

Expected answer components:

- Baseline subgroup fairness metrics (at least demographic parity gap or equal opportunity gap).
- One explicit mitigation intervention (e.g., reweighing) with pre/post comparison.
- Performance-vs-fairness tradeoff analysis.
- Policy constraint check (e.g., max tolerated AUC/F1 degradation).

Minimum acceptable artifacts:

- `q20_fairness_mitigation_comparison.csv`
- `q20_fairness_tradeoff.png`

Grading note: no full credit if mitigation is presented without explicit policy constraints for deployment decisions.

Block J (Bonus): Advanced Extensions

Strong submissions may include:

- **Causal DAG**: clear graph, plausible assumptions, and discussion of (non-)identifiability and adjustment sets.
- **Uncertainty**: conformal prediction or calibrated intervals with empirical coverage reported on held-out data.
- **Temporal validation**: chronological split vs random split with degradation analysis.
- **Online/streaming serving**: feature freshness plan, SLA/latency targets, OOD/drift guardrail, rollback path.

Partial credit for well-reasoned designs even without full code; no credit for causal claims without addressing assumptions.

Capstone

Minimum complete capstone output:

1. leakage-safe preprocessing,
2. best model with validated metrics,
3. SHAP local explanation for high-citation no-migration case,
4. global importance plot,
5. fairness slice and deployment recommendation.

Rubric Notes for TAs

- Deduct for hidden leakage or unjustified assumptions.
- Deduct for non-reproducible code.
- Reward honest limitations and rigorous diagnostics.