UT-ECE Data Science Final Assignment
Complete Solution Manual (Extended & Grading-Oriented)

Teaching Assistant Team

Spring 2025

Purpose of this Manual

This document provides complete conceptual solutions, implementation guidance, and grading-oriented deliverables for Q1–Q6. It is designed so students can map each question to code, report text, and evaluation artifacts.

Contents

Global Assumptions and Reproducibility Standards

- All reported metrics must come from a leakage-safe split (**time-aware if possible**).
- Preprocessing (imputation/scaling/encoding) must be fit on training data only.
- Random seeds should be fixed and reported.
- Post-outcome features are excluded from both model fitting and hyperparameter tuning.
- Statistical claims must include assumptions and uncertainty (CI/p-values/effect sizes when applicable).

Recommended report artifacts per question:

- one concise theory block,
- one implementation block,
- one diagnostics/result block,
- one short "risk & limitation" note.

Q1. Advanced Data Engineering & SQL

Q1A. Window-function solution (complete)

Goal: compute country-level 3-year moving average of citations and rank users by this smoothed signal.

```
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
),
ranked AS (
    SELECT
        UserID,
        Country_Origin,
        Year,
        Research_Citations,
        moving_avg_citations,
        DENSE_RANK() OVER (
            PARTITION BY Country_Origin
            ORDER BY moving_avg_citations DESC
        ) AS country_rank,
        NTILE(10) OVER (
            PARTITION BY Country_Origin
            ORDER BY moving_avg_citations DESC
        ) AS country_decile
    FROM citation_velocity
)
SELECT *
FROM ranked
ORDER BY Country_Origin, country_rank, Year;
```

Why this is correct:

- ROWS BETWEEN 2 PRECEDING AND CURRENT ROW gives a 3-point moving window.
- PARTITION BY Country_Origin prevents cross-country contamination.
- DENSE_RANK handles ties without rank gaps.

Edge-case note: in first two years of each country, moving average is over fewer than 3 observations by design.

Q1B. Leakage diagnosis (complete)

Direct leakage:

- Visa_Approval_Date if timestamp is after decision/event target.

Potential temporal leakage (must verify event time):

- Last_Login_Region
- Passport_Renewal_Status

Usually safe (if measured pre-inference):

- Years_Since_Degree

Leakage audit protocol (recommended):

1. define prediction timestamp t_0,
2. verify each feature timestamp $t_f \le t_0$,
3. drop/lag/aggregate features violating causality order,
4. re-run feature importance to ensure no hidden proxies remain.

Common mistake: checking only semantic meaning of features without checking logged timestamps.

Q2. Statistical Inference & Linear Models

Q2A. Elastic Net gradient and optimization interpretation

Given objective:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \lambda_1\sum_{j=1}^{n}|\theta_j| + \frac{\lambda_2}{2}\sum_{j=1}^{n}\theta_j^2.$$

For coordinate θ_j:

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

Subgradient of absolute value:

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

Implication:

- ℓ_1 term induces sparsity ($\theta_j = 0$ exactly).
- ℓ_2 term stabilizes under collinearity.
- Elastic Net balances feature selection and coefficient shrinkage.

Practical optimizer note:

- coordinate descent is standard for convex EN formulations;
- standardization of features before EN is essential.

Q2B. Coefficient interpretation with uncertainty

Given coefficient 0.52, p-value 0.003, and 95% CI $[0.18, 0.86]$:

- $p < 0.05 \Rightarrow$ reject $H_0 : \beta = 0$.
- CI excludes zero $\Rightarrow$ statistical evidence of non-zero effect.
- Positive interval entirely above 0 $\Rightarrow$ positive association.

For logistic regression: A one-unit increase multiplies odds by $\exp(0.52) \approx 1.68$, ceteris paribus.

Caution:

- significance $\neq$ causal effect;
- coefficient comparability requires feature scaling awareness;
- multicollinearity can inflate uncertainty.

Q3. Optimization & Gradient Descent

Ravine behavior and optimizer comparison

Ravine geometry: steep curvature in one direction, shallow in another. Vanilla SGD oscillates across steep walls and progresses slowly along valley floor.

Momentum dynamics

$$v_t = \beta v_{t-1} + \eta \nabla J(\theta_t), \qquad \theta_{t+1} = \theta_t - v_t.$$

- Opposite-sign gradients in steep axis partially cancel via momentum averaging.
- Consistent gradients in shallow axis accumulate velocity.
- Net effect: reduced zig-zag, faster valley traversal.

Adam dynamics

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \qquad s_t = \beta_2 s_{t-1} + (1-\beta_2) g_t^2.$$

With bias correction:

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{s}_t = \frac{s_t}{1-\beta_2^t}, \quad \theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{s}_t}+\epsilon}.$$

Why Adam often wins in this setup:

- per-coordinate adaptive learning rates,
- robustness to scale heterogeneity,
- less manual learning-rate tuning.

What to submit:

- trajectory plot on contour map,
- loss-vs-iteration plot,
- short recommendation justified by observed curvature behavior.

Q4. Non-Linear Models & Kernels

Q4A. RBF overfitting control

RBF kernel:

$$K(x,x') = \exp(-\gamma \|x-x'\|^2).$$

If overfitting occurs: decrease γ (and/or decrease C).

- Large γ: very local influence $\Rightarrow$ highly flexible boundary.
- Small γ: smoother, broader influence $\Rightarrow$ lower variance.

Hyperparameter interaction:

- High C + high γ: strongest overfit risk.
- Lower C can regularize margin violations.

Q4B. Cost-complexity pruning

$$R_\alpha(T) = R(T) + \alpha |T|.$$

- $\alpha \uparrow \Rightarrow$ stronger penalty on leaf count $\Rightarrow$ smaller tree.
- $\alpha \downarrow \Rightarrow$ larger tree, potentially lower training error but higher variance.

Model-selection protocol:

1. generate pruning path over α,
2. evaluate by CV,
3. choose smallest tree within 1-SE rule (optional, robust practice).

Q5. Unsupervised Learning

Q5A. PCA explained variance ratio

Given covariance eigenvalues $\lambda_1, \lambda_2, \lambda_3$:

$$\mathrm{EVR}(PC_k) = \frac{\lambda_k}{\sum_j \lambda_j}.$$

Interpretation: λ_k is variance captured along principal axis k.

Cumulative criterion:

$$\mathrm{CEVR}(K) = \sum_{k=1}^{K} \mathrm{EVR}(PC_k).$$

Pick minimum K achieving target (e.g., 90%–95%).

Important: PCA should be applied after centering (and typically scaling if units differ).

Q5B. Elbow method for K-Means

Within-cluster sum of squares:

$$\mathrm{WCSS}(K) = \sum_{c=1}^{K} \sum_{x_i \in c} \|x_i - \mu_c\|^2.$$

WCSS decreases monotonically as K increases.

Define marginal gain:

$$\Delta_K = \mathrm{WCSS}(K-1) - \mathrm{WCSS}(K).$$

Elbow is where Δ_K starts shrinking substantially.

Good practice:

- report elbow + silhouette score together;
- run multiple initializations to avoid local minima;
- assess cluster stability under resampling.

Q6. Capstone Explainability

Local SHAP decomposition

For one instance:

- `base_value`: expected model output on background data
- `output_value`: model output for that instance

SHAP additivity:

$$\texttt{output\_value} = \texttt{base\_value} + \sum_{j=1}^{p} \phi_j$$

where ϕ_j is feature j's contribution.

Interpretation example

A high-citation candidate predicted `No Migration` can occur if:

- citation feature contributes positively,
- but multiple stronger negative contributors (e.g., policy-region effects, career-stage patterns, compensation mismatch) dominate total logit/probability shift.

What makes explanation reliable

- background dataset must match deployment population,
- feature pipeline at explanation time must match training/serving pipeline,
- report both local (ϕ_j table/waterfall) and global (mean$|\phi_j|$) views.

Final Deliverables Checklist (Grading-Oriented)

Section	Minimum complete evidence
Q1 SQL & Leakage	Window query output sample, ranking table, timestamp-based leakage audit table (feature, event-time status, action).
Q2 Inference	Elastic Net derivation, coefficient table with CI/p-values, assumptions note.
Q3 Optimization	Trajectory + loss plots for SGD/Momentum/Adam, comparison paragraph.
Q4 Nonlinear	SVM grid/CV heatmap (at least conceptual), pruning path vs validation score.
Q5 Unsupervised	PCA EVR/CEVR plot, elbow+silhouette evidence, cluster interpretation.
Q6 Explainability	One local SHAP case, one global importance plot, consistency and caveat note.

Common Reasons for Point Deduction

- leakage not audited with timestamps,
- reporting metrics without uncertainty or split protocol,
- claiming causal interpretation from observational associations,
- explanation plots without linking to final decision logic,
- missing reproducibility details (seed, versions, split rules).

Short Executive Summary Template

Problem: Predict migration propensity under fairness and reliability constraints.
Approach: Leakage-safe data engineering, regularized supervised models, optimizer diagnostics, nonlinear model control, unsupervised structure analysis, SHAP explainability.
Result: Best model selected via validation protocol with interpretable and auditable behavior.
Deployment note: Proceed with monitoring for drift, calibration, and subgroup fairness.

(Fill with actual numbers):

- Best AUC: `TODO`
- Best F1: `TODO`
- Brier / ECE: `TODO`
- Max subgroup TPR gap: `TODO`