University of Tehran
Department of Electrical and Computer Engineering

Data Science - Final Assignment

Complete Solved Version

Lead Teaching Assistant Team

Spring 2025

Analyzing Global Tech Talent Migration: You are provided with a dataset of 50,000 tech professionals. The goal is to predict `Migration_Status` (1 if they migrated to another country for work, 0 otherwise). Features include `GitHub_Activity`, `Research_Citations`, `Industry_Experience`, and categorical data such as `Education_Level`.

1 Question 1: Advanced Data Engineering & SQL

Part A: Time-Series Trends via Window Functions

Required: Compute 3-year moving average of `Research_Citations`, partitioned by `Country_Origin`, then rank users inside each country by this moving average.

Solution SQL (single-query with CTE):

```
WITH citation_ma AS (
    SELECT
        UserID,
        Country_Origin,
        Year,
        Research_Citations,
        AVG(Research_Citations) OVER (
            PARTITION BY Country_Origin, UserID
            ORDER BY Year
            ROWS BETWEEN 2 PRECEDING AND CURRENT ROW
        ) AS ma3_citations
    FROM Professionals_Data
)
SELECT
    UserID,
    Country_Origin,
    Year,
    Research_Citations,
    ma3_citations,
```

```
    DENSE_RANK() OVER (
        PARTITION BY Country_Origin
        ORDER BY ma3_citations DESC
    ) AS country_rank_by_ma3
FROM citation_ma
ORDER BY Country_Origin, country_rank_by_ma3, UserID, Year;
```

Why this is correct:

- `ROWS BETWEEN 2 PRECEDING AND CURRENT ROW` gives a rolling window of size 3.
- Partition by `Country_Origin, UserID` computes per-user trend inside each country.
- `DENSE_RANK()` ranks users by smoothed citation signal inside each country.

Part B: Diagnostic Identification of Data Leakage

Feature-by-feature diagnosis:

- `Years_Since_Degree`: **Usually safe**, if degree date is known before prediction time.
- `Visa_Approval_Date`: **Direct leakage** (typically post-outcome or tightly post-decision).
- `Last_Login_Region`: **Potential temporal leakage**; if recorded after migration decision, it leaks outcome state.
- `Passport_Renewal_Status`: **Potential leakage**; may reflect post-decision process behavior depending on timestamp.

Rule: a feature is allowed only if its timestamp $t_f \leq t_0$ (prediction time).

2 Question 2: Statistical Inference & Linear Models

Part A: Elastic Net Derivation

Given

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

for coordinate θ_j:

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

Subgradient of $|\theta_j|$:

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

Interpretation at $\theta_j = 0$:

- The derivative is set-valued ($[-1,1]$).
- This enables exact zeros under coordinate descent/proximal updates.
- Hence Elastic Net combines sparsity (L_1) and stability (L_2).

Part B: Interpreting Regression Outputs

Given for `GitHub_Activity`:

$$\hat{\beta} = 0.52, \quad p = 0.003, \quad 95\% \text{ CI} = [0.18, 0.86].$$

Statistical significance under $H_0 : \beta = 0$:

- Since $p = 0.003 < 0.05$, reject H_0.
- CI excludes 0, confirming significance.

Effect interpretation:

- Estimated association is positive.
- If logistic model: one-unit increase multiplies odds by $\exp(0.52) \approx 1.68$.
- If linear probability-style interpretation, sign and CI still indicate positive relationship, with model-specific scale caveat.

3 Question 3: Optimization & Gradient Descent

Ravine phenomenon: In ravine-shaped objectives, curvature is very steep in one direction and flat in another. Plain SGD oscillates across steep walls and advances slowly along the shallow axis.

Momentum:

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

- Reduces zig-zag oscillation by accumulating consistent directional gradients.
- Faster movement along valley floor compared to vanilla SGD.

Adam:

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

with bias correction and per-coordinate scaling

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{s}_t} + \epsilon}.$$

- Typically more robust to anisotropy and feature-scale mismatch.
- Usually converges faster with less manual tuning.

Practical recommendation: Use Adam as default in heterogeneous-scale settings; use Momentum-SGD when you want simpler dynamics and strong control via schedule tuning.

4 Question 4: Non-Linear Models & Kernels

Part A: SVM RBF Parameterization

RBF kernel:

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

If overfitting occurs, **decrease** γ.

- Large γ: very local influence, highly wiggly boundary (high variance).
- Smaller γ: smoother boundary, better generalization.

Also consider reducing C to strengthen regularization.

Part B: Cost-Complexity Pruning

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

- α penalizes tree size ($|T|$: number of leaves/terminal nodes).
- Small α: larger trees (lower bias, higher variance).
- Large α: smaller trees (higher bias, lower variance).

Selection: Choose α via validation/CV to optimize out-of-sample performance rather than training error.

5 Question 5: Unsupervised Learning

Part A: PCA Explained Variance

Let covariance matrix eigenvalues be $\lambda_1, \lambda_2, \lambda_3$ (sorted descending). Explained variance ratio:

$$\text{EVR}(PC_k) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \lambda_3}.$$

Hence:

$$\text{EVR}(PC_1) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}, \qquad \text{EVR}(PC_2) = \frac{\lambda_2}{\lambda_1 + \lambda_2 + \lambda_3}.$$

Interpretation of eigenvalues: λ_k is variance captured along principal axis k. Larger λ_k means more information retained by that component.

Part B: K-Means Elbow Method

WCSS objective:

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

As K increases, WCSS monotonically decreases (more centroids $\Rightarrow$ smaller within-cluster distances).

Define marginal gain:

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

The elbow is where Δ_K begins to shrink sharply, indicating diminishing returns. So elbow gives a geometric complexity-vs-fit compromise.

6 Question 6: Capstone Explainability

For SHAP in XGBoost:

- `base_value`: expected model output over background data.
- `output_value`: prediction for the specific candidate.

  Additive relation:

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

where ϕ_j is feature j's SHAP contribution.

Why high-citation candidate can still be predicted `No Migration`:

- `Research_Citations` may push prediction upward (positive ϕ),
- but stronger negative contributions from other features (e.g., region-policy interaction, stability proxies, experience pattern) can dominate, leading to final negative class.

Important caveat: SHAP explains model behavior, not causal truth.

Compact Grading Checklist

- Q1: correct window function + timestamp-based leakage reasoning
- Q2: correct EN gradient/subgradient + valid statistical interpretation
- Q3: accurate ravine explanation + Momentum vs Adam comparison
- Q4: correct γ direction for overfit + pruning bias-variance logic
- Q5: correct EVR formulas + sound elbow argument
- Q6: correct SHAP decomposition and interpretation