University of Tehran – ECE Department

Data Science Comprehensive Final Assessment (Extended Edition)

Complete Solution Blueprint & Professional Submission Guide

Course Staff (Spring 2025) – Professional Assignment Pack

Version 2.0 (Complete Guide)

Course Context: This extended final integrates the full UT-ECE Data Science track from Python and scientific studies to SQL engineering, ML, deep learning, NLP, LLM agents, and production reliability.

1. Assessment Overview

Primary dataset: `GlobalTechTalent_50k.csv` (50,000 rows).
Primary target: `Migration_Status` (binary).

Additional allowed datasets:

- Course assignment datasets from UT-ECE repositories (where relevant).
- Public benchmark datasets with proper citation.

Required submission artifacts:

1. One reproducible notebook with clear sectioning (Q1–Q20 + Capstone).
2. One PDF report (max 20 pages excluding appendix).
3. One code package with scripts/modules and dependency file.
4. One presentation deck (10–15 slides).
5. One ethics/fairness memo (1–2 pages).

Reproducibility requirements:

- Fix random seeds where applicable.
- State train/validation/test splitting strategy.
- Log software environment and package versions.
- No leakage from post-outcome variables.

Recommended execution order: Setup → Leakage Audit → EDA/Inference → Supervised/Unsupervised models → Deep/NLP → Fairness/Governance → Capstone integration → Q15–Q20 production diagnostics.

2. Grading Distribution (260 points)

Block	Focus	Points
A	Foundations: lifecycle, Python, EDA, scientific studies	20
B	Inference + visualization design and storytelling	20
C	SQL engineering + big-data systems thinking	25
D	Supervised ML + optimization + model selection	45
E	Unsupervised learning + dimensionality reduction	20
F	Deep learning + NLP + LMs/LLM agents	30
G	Ethics, fairness, robustness, governance	15
H	Integrated capstone implementation + communication	25
I	Production reliability and advanced diagnostics (Q15–Q20)	60
J (Bonus)	Advanced research/production extensions (optional)	+20
Total (A–I)		260
Optional bonus		+20

Block A – Foundations (20 points)

Q1. Data Science Lifecycle and Problem Framing (10 pts)

Expected complete answer:

- **Business objective:** predict migration propensity for policy/retention decisions.
- **Success criteria:** define AUC/F1/calibration/fairness targets.
- **Assumptions:** timestamp correctness, label validity, representativeness.
- **Failure modes:** leakage, sampling bias, drift, proxy discrimination.
- **Deployment:** batch or API scoring + human review policy.
- **Monitoring:** performance, calibration, fairness, drift thresholds.

Deliverable format: one-page structured statement + lifecycle diagram:

$$\text{Problem} \rightarrow \text{Data} \rightarrow \text{Model} \rightarrow \text{Eval} \rightarrow \text{Deploy} \rightarrow \text{Monitor} \rightarrow \text{Retrain}$$

Q2. Python Data Operations and EDA (10 pts)

Minimum complete implementation:

1. Schema checks: dtypes, nulls, duplicates, outliers, invalid ranges.
2. At least 6 plots:
   - target balance,
   - missingness profile,
   - distributions of key numeric variables,
   - boxplots by target,
   - correlation heatmap,
   - group migration rate (country/education).
3. One reusable preprocessing function + tests.

Suggested utility signature:

```
def build_preprocessor(num_cols, cat_cols):
    # impute/scale numeric, impute/encode categorical
    # return sklearn ColumnTransformer pipeline
    ...
```

Unit tests required: no NaN after transform, stable output shape, unseen-category handling.

Block B – Inference and Visualization (20 points)

Q3. Scientific Studies and Inference (10 pts)

Expected complete answer:

- Observational vs experimental framing (causal limits).
- Sampling-bias risks and mitigation.
- One confidence interval + one hypothesis test with assumptions.

Example CI (difference in two proportions):

$$(\hat{p}_1 - \hat{p}_2) \pm z_{0.975}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example hypothesis test: chi-square independence (education vs migration) with effect size (Cramérs V).

Q4. Visualization Design + Storytelling (10 pts)

Required elements in dashboard/narrative:

- KPI definitions (migration rate, risk count, threshold metrics, fairness gap).
- Preattentive and color rationale (consistency, contrast, cognitive load).
- One misleading chart pitfall + corrected version.

Pitfall example: truncated y-axis exaggerating group differences. **Fix:** proper baseline, confidence bars, annotation.

Block C – SQL and Data Engineering (25 points)

Q5. SQL-1/SQL-2 Advanced Querying (15 pts)

(i) 3-year moving average by country (window):

```
SELECT
  country_origin,
  year,
  AVG(research_citations) OVER (
    PARTITION BY country_origin
    ORDER BY year
    ROWS BETWEEN 2 PRECEDING AND CURRENT ROW
  ) AS ma3_citations
FROM professionals_data;
```

(ii) top decile + percentile bucketing:

```
SELECT
  userid,
  research_citations,
  NTILE(10) OVER (ORDER BY research_citations DESC) AS decile,
  PERCENT_RANK() OVER (ORDER BY research_citations) AS pct_rank
FROM professionals_data;
```

(iii) cohort retention (CTE style):

```
WITH base AS (
  SELECT cohort_year, migration_status
  FROM candidate_outcomes
),
cohort_size AS (
  SELECT cohort_year, COUNT(*) AS n_total
  FROM base GROUP BY cohort_year
),
retained AS (
  SELECT cohort_year, COUNT(*) AS n_not_migrated
  FROM base
  WHERE migration_status = 0
  GROUP BY cohort_year
)
SELECT c.cohort_year,
       c.n_total,
       r.n_not_migrated,
       1.0 * r.n_not_migrated / c.n_total AS retention_rate
FROM cohort_size c
JOIN retained r USING (cohort_year)
ORDER BY c.cohort_year;
```

Q6. Data Leakage and Big-Data Architecture (10 pts)

Expected complete answer:

- Identify leaky/post-outcome features by timestamp logic $t_f \leq t_0$.
- Propose Bronze/Silver/Gold (or equivalent) architecture.
- Explain feature store split:
  - offline store (training; point-in-time correct),
  - online store (serving; low latency),
  - shared feature definitions/versioning.

Block D – Supervised Learning and Optimization (45 points)

Q7. Linear/Logistic Models + Regularization (15 pts)

Minimum complete coverage:

- baseline linear/logistic models,
- Elastic Net objective and gradient/subgradient explanation,
- coefficient interpretation + CI/p-values (when applicable),

- probability calibration check.

Elastic Net form:

$$\mathcal{L}(\beta) + \lambda \left[\alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2 \right]$$

Q8. Optimization Deep Dive (10 pts)

Compare SGD, Momentum, Adam on ravine objective.

Expected observations:

- SGD: oscillatory in steep direction.
- Momentum: damped oscillation, faster valley traversal.
- Adam: adaptive coordinate-wise scaling, robust under heterogeneous scales.

Required artifacts: trajectory plot + loss-vs-iteration + recommendation.

Q9. Model Family Comparison (20 pts)

Model families: SVM/KNN, Tree/RF, one boosting model.

Required protocol:

1. Cross-validation setup (stratified or temporal).
2. Hyperparameter search (random/grid/Bayesian with bounds).
3. Error analysis: confusion patterns + subgroup slices.

Minimum report table:

Model	AUC	F1	Precision	Recall
Logistic	–	–	–	–
RandomForest	–	–	–	–
Boosting	–	–	–	–

Block E – Unsupervised Learning (20 points)

Q10. Dimensionality Reduction (10 pts)

Required:

- PCA explained variance ratio (EVR, cumulative EVR plot),
- one additional method (RP/t-SNE/UMAP),
- interpretation limits of latent dimensions.

Q11. Clustering (10 pts)

Required:

- K-Means + elbow and silhouette,
- DBSCAN (or density equivalent),
- cluster stability and practical meaning.

Note: justify chosen K/ϵ and discuss sensitivity.

Block F – Deep Learning, NLP, and LMs (30 points)

Q12. Neural Networks and Sequence Models (15 pts)

Required experiments:

- one tabular NN (MLP/shallow FFN),
- one sequence/NLP model (CNN/RNN/LSTM/GRU),
- comparison vs best classical baseline.

Must report: split protocol, early stopping logic, overfitting diagnostics.

Q13. Language Models and LLM Agents (15 pts)

Expected complete answer:

- agentic workflow (retrieve → plan → tool use → verify → respond),
- evaluation criteria: faithfulness, hallucination rate, safety,
- governance constraints: PII controls, prompt-injection defenses, audit logs.

Block G – Ethics and Governance (15 points)

Q14. Fairness, Bias, and Responsible Deployment (15 pts)

Required:

- subgroup metrics (country/education etc.),
- discussion of historical bias + proxy discrimination,
- human-in-the-loop, override and appeals policy.

Recommended fairness metrics: TPR/FPR gaps, precision parity, calibration by subgroup, DP gap (if policy-relevant).

Block H – Integrated Capstone (25 points)

Capstone Task

Implementation must include:

1. leakage-safe preprocessing and training pipeline,
2. model card + experiment tracking summary,
3. SHAP local and global explainability,
4. deployment recommendation + monitoring thresholds.

Mandatory outputs:

- one local explanation for a high-citation candidate predicted no-migration,
- one global feature-importance plot,
- one fairness slice table,
- one executive summary for non-technical stakeholders.

Block I – Production Reliability Extension (60 points)

Q15. Calibration and Threshold Policy (10 pts)

Required:

- reliability/calibration curve,
- at least one calibration metric (Brier, ECE),
- threshold maximizing F1,
- threshold minimizing asymmetric cost ($C_{FN} > C_{FP}$).

Deliverables: calibration plot, threshold tradeoff plot, final threshold recommendation with policy logic.

Q16. Drift Detection and Monitoring Design (10 pts)

Required:

- PSI ranking for numeric features,
- one categorical drift metric (e.g., JS divergence),
- warning/critical thresholds and retraining triggers.

SOP expectation: who monitors, frequency, escalation policy, rollback/retrain condition.

Q17. Counterfactual Recourse Analysis (10 pts)

Required:

- select at least two actionable features,
- minimal intervention under feasibility caps,
- recourse success rate + median intervention by feature.

Include ethics note: avoid unrealistic/inequitable recourse recommendations.

Q18. Temporal Backtesting and Rolling Validation (10 pts)

Required:

- chronological folds (or justified fallback ordering),
- fold-wise AUC/F1 and decay vs first fold,
- drift-aware interpretation (e.g., mean PSI trend).

Deliverables: `q18_temporal_backtest.csv`, degradation figure, fallback explanation in report.

Q19. Uncertainty Quantification and Coverage (10 pts)

Required:

- conformal or calibrated predictive intervals/sets,
- empirical coverage across confidence levels,
- interval width and under-coverage analysis.

Deliverables: `q19_coverage_summary.csv`, coverage-vs-confidence figure, low-confidence handling policy.

Q20. Fairness Mitigation Experiment (10 pts)

Required:

- baseline fairness metrics,
- one mitigation (reweighing/thresholding/justified method),
- pre/post fairness and utility comparison,
- explicit policy constraint check (e.g., max AUC drop).

Deliverables: `q20_mitigation_comparison.csv`, fairness-performance tradeoff figure, deployment recommendation.

Block J – Advanced Extensions (Bonus +20)

Any subset earns partial bonus. Keep results reproducible and justified.

1. **Causal framing (5 pts):** DAG, valid/invalid adjustment sets, identifiability limits.
2. **Uncertainty (5 pts):** conformal/calibrated intervals with empirical coverage.
3. **Temporal robustness (5 pts):** time-based vs random split comparison with drift-aware degradation.
4. **Streaming/online serving (5 pts):** minimal online inference design (freshness, idempotence, SLA) + OOD guardrail.

3. Academic Integrity and Professional Standards

- Cite all external resources and model-generated assistance.
- Any copied code without attribution is a violation.
- Report negative results honestly.
- Prefer interpretable, audited pipelines over leaderboard-only optimization.

4. Extra Bonus (up to +10 points)

- Causal inference extension (DAG + identification discussion).
- Real-time pipeline prototype for streaming updates.
- Advanced uncertainty quantification (conformal or Bayesian approximation).

5. Submission Quality Checklist (Must Pass)

All sections Q1–Q20 + Capstone present and clearly labeled.

Leakage audit documented with timestamp logic.

Reproducibility: seeds, split protocol, package versions.

Calibration, drift, uncertainty, fairness all included.

Policy thresholds and governance decisions justified.

Figures/tables filenames consistent with notebook outputs.

Final note: Full credit requires not only strong metrics, but also methodological validity, interpretability, fairness accountability, and production-readiness.