University of Tehran – ECE Department

Data Science Comprehensive Final Assessment (Extended Edition)

Complete Professional Solution Pack

Student Name: ______________________________ Student ID: _____________________

Spring 2025 – Version 1.0

Submission Note

This report is a complete end-to-end solution template for Q1–Q20 + Capstone. Replace all placeholders (`TODO`) with your actual computed values, tables, and figures from the notebook.

Contents

1 Assessment Overview and Reproducibility Protocol

1.1 Dataset and Target

Primary dataset: `GlobalTechTalent_50k.csv` (50,000 rows)
Target: `Migration_Status` (binary: 0/1)

1.2 Submitted Artifacts

1. Reproducible notebook with sections Q1–Q20 + Capstone
2. PDF report (this file)
3. Code package (`src/`, `tests/`, `requirements.txt`)
4. Presentation deck (10–15 slides)
5. Ethics/Fairness memo (1–2 pages)

1.3 Reproducibility Settings

- Global random seed fixed: `TODO_SEED`
- Data split strategy: `TODO_SPLIT_STRATEGY`
- Environment logging: Python `TODO_PY_VERSION`, packages exported
- Leakage controls: all post-outcome features removed prior to training

1.4 Evaluation Protocol

- Validation method: `TODO_CV_METHOD`
- Primary metrics: AUC, F1, Precision, Recall
- Reliability metrics: Brier Score, ECE
- Fairness metrics: subgroup TPR/FPR/Precision gaps, calibration by group

2 Block A – Foundations (20 points)

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

Business Objective

Predict candidate migration propensity to support evidence-based interventions in talent retention and policy planning.

Measurable Success Criteria

- Model discrimination: AUC $\geq$ `TODO_AUC_TARGET`
- Operational utility: F1 at selected threshold $\geq$ `TODO_F1_TARGET`
- Risk control: false negative rate in high-priority segment $\leq$ `TODO`
- Fairness: maximum subgroup TPR gap $\leq$ `TODO_FAIRNESS_BOUND`

Data Assumptions

- Features are measured prior to migration outcome timestamp.
- Labels are sufficiently accurate and consistent.
- Sample is reasonably representative of deployment population.

Potential Failure Modes

- Sampling bias by geography/field
- Label leakage via post-outcome variables
- Temporal drift in macroeconomic/policy conditions
- Proxy discrimination through correlated features

Deployment and Monitoring

- Batch scoring cadence: `TODO_WEEKLY/MONTHLY`
- Human review for low-confidence and policy-sensitive cases
- Automated alerts for drift, calibration decay, and fairness degradation

Lifecycle Diagram

Problem Framing → Data Ingestion → Validation/Cleaning → Feature Engineering → Modeling → Evaluation → Deployment → Monitoring/Retraining

Figure 1: End-to-end lifecycle for migration prediction

2.2 Q2. Python Data Operations and EDA (10 pts)

Schema and Quality Checks

Performed robust checks for:

- Data types, missingness, duplicates, invalid domain values
- Numeric outliers (IQR/Z-score)
- Category normalization and rare category handling

EDA Visuals and Interpretation

At least six plots were produced:

1. Class distribution of `Migration_Status`
2. Missingness profile
3. Core numeric distributions
4. Boxplots by target class
5. Correlation matrix of numeric features

6. Migration rate by country/education

Key finding summary: `TODO_EDA_SUMMARY`

Reusable Preprocessing Utility

Implemented function: `build_preprocessor(...)` with:

- Numeric branch: imputation + scaling
- Categorical branch: imputation + one-hot encoding
- Unseen category handling for validation/test/inference

Unit Tests

- Output shape invariance
- No nulls after transform
- Stable behavior under fixed seed
- Correct behavior on unseen categories

3 Block B – Inference and Visualization (20 points)

3.1 Q3. Scientific Studies and Inference (10 pts)

Observational vs Experimental

This study is observational; causal conclusions are limited without randomization or valid identification assumptions.

Sampling Bias Risks

- Over/under representation across countries and institutions
- Platform visibility bias
- Survivorship bias in profile availability

Confidence Interval (Example)

For migration rate difference between two groups:

$$(\hat{p}_1 - \hat{p}_2) \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Computed result: `TODO_CI_RESULT`

Hypothesis Test (Example)

- H_0: migration is independent of education level
- Test: Chi-square test of independence
- Reported: $\chi^2 =$ `TODO`, $p =$ `TODO`, effect size (Cramérs V) = `TODO`

3.2 Q4. Visualization Design + Storytelling (10 pts)

Stakeholder KPI Dashboard

Included:

- Overall migration rate
- Predicted high-risk count
- Threshold-specific Precision/Recall/F1
- Fairness gap indicator

Design Rationale

- Consistent color mapping for favorable/unfavorable outcomes
- Preattentive emphasis on position/length before color
- Uncertainty bars for subgroup comparisons

Misleading Visualization Pitfall and Fix

- Pitfall: truncated y-axis exaggerating subgroup gaps
- Correction: baseline-aware axes + clear annotation + CI bars

4 Block C – SQL and Data Engineering (25 points)

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

(1) 3-year moving average citations by country

```
SELECT
  country,
  year,
  AVG(avg_citations) OVER (
    PARTITION BY country
    ORDER BY year
    ROWS BETWEEN 2 PRECEDING AND CURRENT ROW
  ) AS ma3_citations
FROM country_year_stats;
```

(2) Top decile ranking + percentile bucketing

```
SELECT
  candidate_id,
  total_citations,
  NTILE(10) OVER (ORDER BY total_citations DESC) AS decile,
  PERCENT_RANK() OVER (ORDER BY total_citations) AS pct_rank
FROM candidates;
```

(3) Cohort retention (CTE style)

```
WITH base AS (
  SELECT candidate_id, cohort_year, migration_status
  FROM candidate_outcomes
),
cohort_size AS (
  SELECT cohort_year, COUNT(*) AS n0
  FROM base GROUP BY cohort_year
),
retained AS (
  SELECT cohort_year, COUNT(*) AS n_retained
  FROM base
  WHERE migration_status = 0
  GROUP BY cohort_year
)
SELECT
  c.cohort_year, c.n0, r.n_retained,
  1.0 * r.n_retained / c.n0 AS retention_rate
FROM cohort_size c
JOIN retained r USING (cohort_year)
ORDER BY c.cohort_year;
```

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

Leaky Feature Audit

Excluded features containing post-outcome information:

- `TODO_LEAKY_FEATURE_1`
- `TODO_LEAKY_FEATURE_2`
- `TODO_LEAKY_FEATURE_3`

Batch + Streaming Architecture

- Bronze: raw immutable ingestion (batch + stream)
- Silver: validated, standardized, deduplicated data
- Gold: model-ready feature tables and KPI marts

Feature Store Design

- Offline store for training (point-in-time correct)
- Online store for low-latency serving
- Unified feature definitions and versioning to ensure train/serve parity

5 Block D – Supervised Learning and Optimization (45 points)

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

Baselines

- Baseline logistic regression for binary outcome

- Optional linear baseline for continuous proxy task (if applicable)

Elastic Net Objective

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

Implementation approach: `TODO_LIBRARY_OR_CUSTOM_GRADIENT`

Interpretation

- Coefficient sign and magnitude interpretation in log-odds
- Confidence intervals (where inferential framework available)
- Calibration check on predicted probabilities

5.2 Q8. Optimization Deep Dive (10 pts)

Compared SGD, Momentum, Adam on ravine objective:

$$f(x, y) = 100(y - x^2)^2 + (1 - x)^2$$

Observed behavior:

- SGD: oscillatory in steep curvature direction
- Momentum: smoother convergence, reduced oscillation
- Adam: robust and fast under feature-scale heterogeneity

Recommendation: `TODO_OPTIMIZER_RECOMMENDATION`

5.3 Q9. Model Family Comparison (20 pts)

Model Set

- SVM, KNN
- Decision Tree, Random Forest
- Boosting model (XGBoost/GBM/CatBoost)

Search Protocol

- Cross-validation: `TODO_CV`
- Hyperparameter tuning: `TODO_RANDOM/BAYESIAN/GRID`
- Final lock: best validation model evaluated once on test

Performance Summary

Table 1: Model comparison on validation/test

Model	AUC	F1	Precision	Recall
Logistic Regression	TODO	TODO	TODO	TODO
Random Forest	TODO	TODO	TODO	TODO
Boosting Model	TODO	TODO	TODO	TODO
Best Model (TODO)	TODO	TODO	TODO	TODO

Error Analysis

Confusion patterns indicated TODO_ERROR_PATTERN. Additional analysis by subgroup suggests TODO_SUBGROUP_RISK.

6 Block E – Unsupervised Learning (20 points)

6.1 Q10. Dimensionality Reduction (10 pts)

PCA

- Cumulative explained variance at k components: TODO
- Selected dimension: TODO_K

Additional Method

Used TODO_TSNE/UMAP/RP for latent structure visualization/embedding.
Interpretation: TODO_LATENT_INTERPRETATION

6.2 Q11. Clustering (10 pts)

K-Means

- Elbow-selected k: TODO
- Silhouette score: TODO

DBSCAN

- Parameters: $\epsilon =$ TODO, min_samples=TODO
- Noise rate: TODO

Stability and Meaning

Cluster stability across seeds/resamples: TODO. Practical profile summary: TODO.

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

7.1 Q12. Neural Networks and Sequence Models (15 pts)

Tabular NN

MLP architecture: TODO_ARCH with dropout/batchnorm and early stopping.

Sequence/NLP Model

Model: `TODO_LSTM/GRU/CNN` on text/sequence feature variant. Tokenization and max length: `TODO`.

Comparison Against Classical Baseline

Table 2: Deep model vs classical baseline

Model	AUC	F1	Notes
Best Classical (`TODO`)	`TODO`	`TODO`	`TODO`
MLP	`TODO`	`TODO`	`TODO`
Sequence/NLP	`TODO`	`TODO`	`TODO`

7.2 Q13. Language Models and LLM Agents (15 pts)

Agentic Workflow (Design)

1. Retrieve grounded knowledge (policy/docs/DB rows)
2. Plan task decomposition
3. Tool execution (query/scoring)
4. Compose citation-grounded response
5. Safety and policy filter

Evaluation Criteria

- Faithfulness to retrieved context
- Hallucination rate
- Safety violation rate
- Latency and cost per request

Governance Constraints

- PII minimization and access controls
- Prompt injection defenses
- Audit logging and human escalation path

8 Block G – Ethics, Fairness, Governance (15 points)

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

Subgroup Evaluation

Evaluated by country/education (and additional relevant groups):

- TPR/FPR gaps
- Precision disparities

- Calibration differences

Table 3: Fairness slice summary (baseline)

Group	TPR	FPR	Precision	Support
Group A	`TODO`	`TODO`	`TODO`	`TODO`
Group B	`TODO`	`TODO`	`TODO`	`TODO`
Group C	`TODO`	`TODO`	`TODO`	`TODO`

Bias Discussion

Potential historical-policy and representation biases identified: `TODO`.

Human-in-the-Loop Policy

- Manual override for low-confidence/high-impact predictions
- Appeals process with documented rationale
- Periodic fairness and impact audit

9 Block H – Integrated Capstone (25 points)

9.1 Capstone Implementation Summary

1. Leakage-safe preprocessing and model training pipeline
2. Model card and experiment tracking summary
3. SHAP global and local explainability outputs
4. Deployment recommendation with thresholds and monitoring

9.2 Required Capstone Outputs

(1) Local Explanation

Case: high-citation candidate predicted as no-migration.
Top feature contributions: `TODO_LOCAL_SHAP_SUMMARY`

(2) Global Feature Importance

Global SHAP ranking indicates most influential features: `TODO`.

(3) Fairness Slice Table

Included in Section Q14 and updated for final selected threshold.

(4) Executive Summary

Executive Summary for Non-Technical Stakeholders

The final model improves migration risk identification over baseline methods while maintaining transparent decision logic and fairness checks. The recommended threshold balances overall accuracy and policy cost asymmetry. Deployment is recommended with guardrails: continuous drift monitoring, calibration checks, and mandatory human review for low-confidence or high-impact cases.

10 Block I – Production Reliability Extension (Q15–Q20, 60 points)

10.1 Q15. Calibration and Threshold Policy (10 pts)

Calibration Analysis

- Reliability curve generated
- Brier Score: `TODO`
- ECE: `TODO`

Threshold Policies

- Threshold maximizing F1: `TODO`
- Threshold minimizing asymmetric cost ($C_{FN} > C_{FP}$): `TODO`

Final Threshold Recommendation

`TODO_THRESHOLD_JUSTIFICATION`

10.2 Q16. Drift Detection and Monitoring Design (10 pts)

Drift Metrics

- PSI for numeric features
- JS divergence for categorical distribution shift

Table 4: Drift table (window A vs window B)

Feature	Metric	Value	Status
Feature_1	PSI	TODO	TODO
Feature_2	PSI	TODO	TODO
Country Dist.	JS Div.	TODO	TODO

Monitoring SOP

- Warning threshold: `TODO`
- Critical threshold: `TODO`
- Retraining trigger: `TODO_TRIGGER_RULE`

10.3 Q17. Counterfactual Recourse Analysis (10 pts)

Setup

Analyzed near-boundary negatives and selected actionable features:

- TODO_ACTIONABLE_1
- TODO_ACTIONABLE_2

Results

- Recourse success rate: TODO
- Median intervention magnitude: TODO

Table 5: Counterfactual recourse examples

Candidate	Feature Change	Required Delta	Outcome Flip
ID_1	TODO	TODO	Yes/No
ID_2	TODO	TODO	Yes/No
ID_3	TODO	TODO	Yes/No

Ethics/Practicality

TODO_RECOURSE_ETHICS_DISCUSSION

10.4 Q18. Temporal Backtesting and Rolling Validation (10 pts)

Method

Chronological folds were used based on TODO_TIME_COLUMN. If unavailable, fallback ordering strategy: TODO_FALLBACK.

Table 6: Temporal backtest summary

Fold	AUC	F1	Decay vs Fold 1
Fold 1	TODO	TODO	0.00
Fold 2	TODO	TODO	TODO
Fold 3	TODO	TODO	TODO

Drift-Aware Interpretation

Performance decay aligned with drift proxy (mean PSI = TODO): TODO_INTERPRETATION.

10.5 Q19. Uncertainty Quantification and Coverage (10 pts)

Method

Implemented TODO_CONFORMAL/CALIBRATED_INTERVALS on validation-calibrated predictions.

Table 7: Coverage summary across confidence levels

Confidence Level	Empirical Coverage	Avg Interval Width
80%	TODO	TODO
90%	TODO	TODO
95%	TODO	TODO

Policy for Low Confidence

TODO_LOW_CONFIDENCE_ESCALATION_RULE

10.6 Q20. Fairness Mitigation Experiment (10 pts)

Baseline vs Mitigated

Mitigation method: TODO_REWEIGHING/THRESHOLDING/OTHER

Table 8: Pre/post mitigation utility-fairness comparison

Metric	Baseline	Mitigated
AUC	TODO	TODO
F1	TODO	TODO
TPR Gap	TODO	TODO
FPR Gap	TODO	TODO

Policy Constraint Evaluation

Constraint example: max AUC drop ≤ 0.02.
Observed change: TODO.
Deployment recommendation: TODO_GO/NO-GO/CONDITIONAL

11 Block J – Advanced Extensions (Bonus +20)

Extension 1: Causal Framing (DAG)

TODO_DAG_DESCRIPTION_AND_IDENTIFICATION_LIMITS

Extension 2: Advanced Uncertainty

TODO_CONFORMAL_EXTENSION_RESULTS

Extension 3: Temporal Robustness

TODO_RANDOM_VS_TEMPORAL_COMPARISON

Extension 4: Streaming/Online Serving

TODO_ONLINE_INFERENCE_DESIGN_WITH_OOD_GUARDRAILS

12 Academic Integrity and Professional Standards

- All external resources, packages, and generated assistance are cited.
- No unattributed copied code.
- Negative or null results are reported transparently.
- Preference given to interpretable and auditable pipelines over leaderboard-only optimization.

13 Conclusion

This project delivers a complete, reproducible, and governance-aware migration prediction pipeline. Final recommendation is based on joint optimization of utility, calibration, fairness, and operational reliability rather than a single metric. Production release is conditionally approved with continuous monitoring, retraining triggers, and human oversight safeguards.

A Appendix A: Figure Placeholders

- Calibration curve: `figures/q15_calibration.png`
- Threshold tradeoff: `figures/q15_threshold_tradeoff.png`
- Drift ranking: `figures/q16_drift_ranking.png`
- Recourse effort: `figures/q17_recourse_effort.png`
- Temporal degradation: `figures/q18_degradation.png`
- Coverage vs confidence: `figures/q19_coverage.png`
- Fairness-utility tradeoff: `figures/q20_tradeoff.png`

B Appendix B: CSV Deliverables

- `outputs/q18_temporal_backtest.csv`
- `outputs/q19_coverage_summary.csv`
- `outputs/q20_mitigation_comparison.csv`

C Appendix C: Environment

```
TODO_PASTE_requirements.txt_OR_conda_env
```