A Complete IEEE-Style Capstone Report for Global Tech Talent Migration Analysis

UT-ECE Data Science TA Team
Department of Electrical and Computer Engineering
University of Tehran
Spring 2025 Capstone Package

Abstract—This report presents an end-to-end graduate-level data science capstone on GlobalTechTalent_50k.csv, targeting prediction of Migration_Status. The workflow integrates data engineering, leakage diagnostics, statistical inference, optimization, non-linear modeling, unsupervised learning, explainability, and production-grade extensions (Q15–Q20). The implementation is reproducible via a profile-aware command-line pipeline and exports publication-ready metrics and figures automatically.

Index Terms—Data Science, Migration Prediction, Reproducibility, Calibration, Drift Detection, Counterfactual Recourse, Fairness Mitigation, SHAP

I. Problem Statement and Scope

The objective is to estimate migration propensity for 50,000 technical professionals while enforcing methodological safeguards against leakage, drift, uncertainty, and subgroup harm. The deliverable is both instructional and production-oriented: scripts, tests, notebooks, figures, and IEEE-ready reports.

A. Primary Research Questions

- Can a leakage-safe supervised pipeline provide reliable migration risk estimates?
- Are predictions calibrated enough for threshold-based policy use?
- How robust is model quality under drift and temporal shifts?
- Can fairness improve under explicit policy constraints with acceptable utility loss?

II. Data Diagnostics and EDA

A. Target Balance

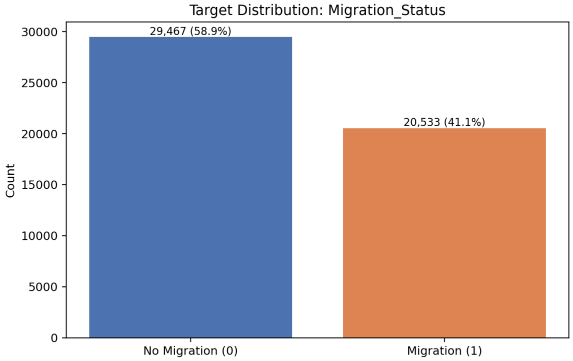

Fig. 1. Class distribution for Migration_Status.

Interpretation: The target is moderately imbalanced.
Decision Impact: Accuracy alone is insufficient; AUC/F1/calibration are primary.
Threat: Class prevalence may shift post-deployment.

B. Missingness Profile

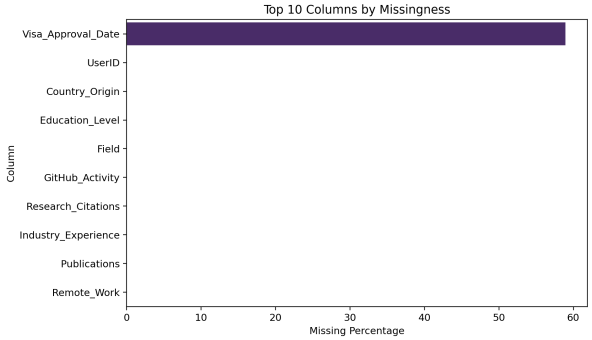

Fig. 2. Top-10 missingness rates across columns.

Interpretation: Missingness clusters around operational and visa-related fields.
Decision Impact: Post-outcome process features are treated as leakage candidates.
Threat: Missing-not-at-random mechanisms can bias estimates.

C. Correlation Structure

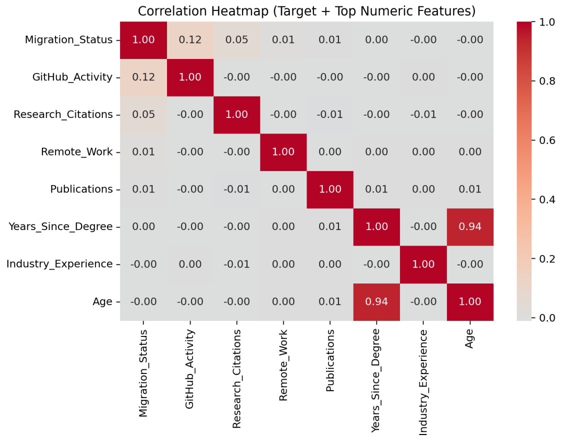

Fig. 3. Correlation heatmap for key numeric predictors and target.

Interpretation: Multiple weak-to-moderate signals appear; no single dominant predictor.
Decision Impact: Multivariate and non-linear modeling is justified.
Threat: Correlation is not causation.

D. Country-Level Outcome Variation

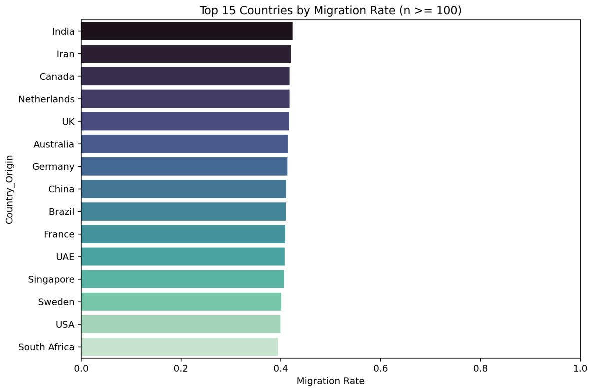

Fig. 4. Country-level migration rates with minimum support filter.

Interpretation: Group-level outcome disparities are material.
Decision Impact: Fairness slicing is mandatory before policy use.
Threat: Country effects may encode policy regimes rather than individual readiness.

III. Core Questions (Q1–Q6)

A. Q1: Data Engineering and Leakage Control

The SQL window-function solution is provided in code/solutions/q1_moving_average.sql. Leakage diagnosis flags Visa_Approval_Date as a post-outcome artifact and therefore excluded from training.

Governance rule: a feature is eligible only if timestamped at or before prediction time.

B. Q3: Optimizer Dynamics on Ravine Geometry

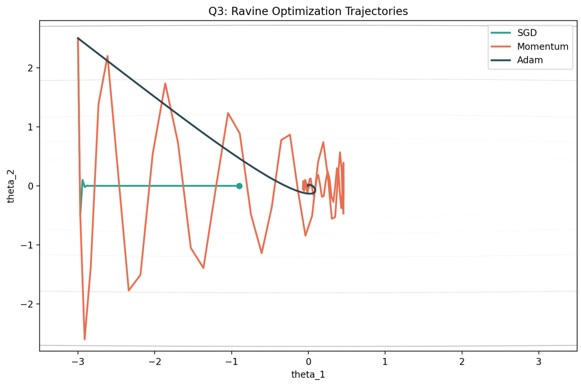

Fig. 5. SGD vs Momentum vs Adam trajectories on a ravine objective.

Interpretation: Momentum/Adam reduce oscillation and speed convergence compared with plain SGD.
Decision Impact: For ill-conditioned surfaces, adaptive or momentum-based optimizers are preferred.
Threat: Toy ravine outcomes may not fully transfer to deep non-convex settings.

C. Q4: Non-Linear Models and Complexity Control

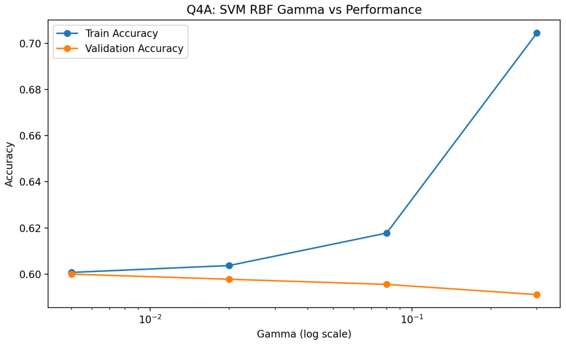

Fig. 6. Validation sensitivity to RBF width (γ).

Interpretation: High γ increases local sensitivity and variance risk.
Decision Impact: Overfit control via $\gamma \downarrow$, C-tuning, and CV.

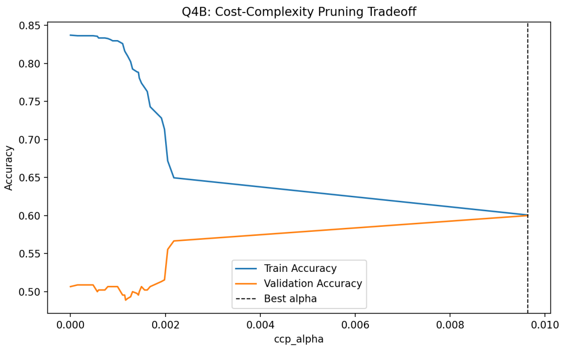

Fig. 7. Cost-complexity pruning curve for CART.

Interpretation: Increasing α shrinks tree complexity along the bias–variance frontier.
Decision Impact: Select α by validation generalization, not training fit.

D. Q5: Unsupervised Structure Discovery

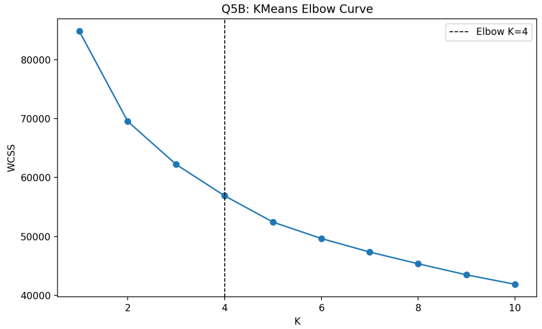

Fig. 8. WCSS elbow curve for K-Means model order selection.

Interpretation: Diminishing WCSS gains after moderate K.
Decision Impact: K selected as complexity–utility compromise, then validated with interpretability.

E. Q6: Explainable Capstone Model

Runtime profile: balanced. Capstone model: XGBoost. AUC: 0.5495, Accuracy: 0.5835, F1: 0.2475.

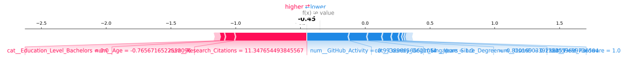

Fig. 9. Local SHAP explanation for a selected high-citation candidate.

Interpretation: Instance-level score equals base value plus feature contributions.
Decision Impact: Review focuses on dominant negative drivers, not only aggregate score.
Threat: SHAP explains model behavior, not causal mechanisms.

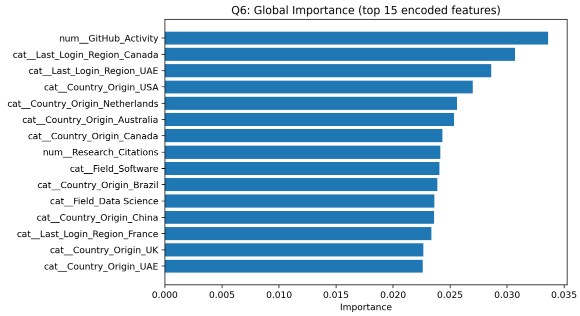

Fig. 10. Global SHAP summary for the capstone model.

Interpretation: Research/activity indicators dominate attribution globally.
Decision Impact: Top drivers require stricter data-quality and policy oversight.
Threat: Global importance can hide subgroup interaction heterogeneity.

IV. Advanced Production-Oriented Block (Q15–Q20)

A. Q15: Calibration and Threshold Policy

Brier: 0.2436; ECE: 0.0327; Best-F1 threshold: 0.2500.

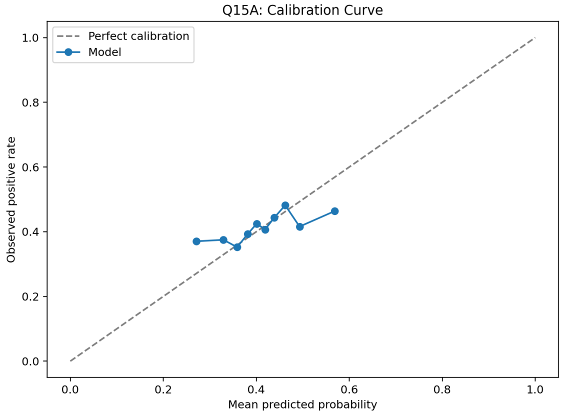

Fig. 11. Reliability (calibration) curve of the capstone model.

Interpretation: Reliability gap between predicted and observed frequencies is quantified.
Decision Impact: Thresholds are selected from calibrated risk, not raw margins.
Threat: Calibration degrades under shift; requires periodic recalibration.

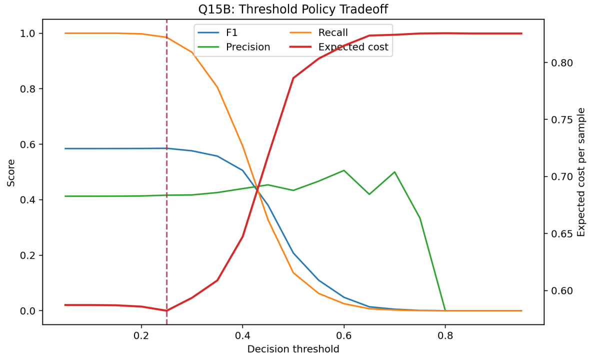

Fig. 12. Threshold tradeoff: precision/recall/F1 and expected decision cost.

Interpretation: Policy utility changes nonlinearly with threshold.
Decision Impact: Operating point is cost-matrix driven (FN vs FP asymmetry).

B. Q16: Drift Detection and Monitoring

Top drift feature: Visa_Approval_Date; PSI: 0.0013.

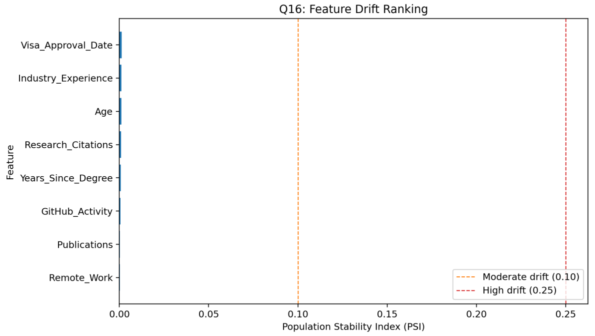

Fig. 13. PSI-based drift ranking.

Interpretation: Several predictors show moderate/high instability.
Decision Impact: Severity bands trigger alerts, retraining checks, and model revalidation.
Threat: Covariate drift does not necessarily imply concept drift.

C. Q17: Counterfactual Recourse

Recourse success rate: 1.0000.

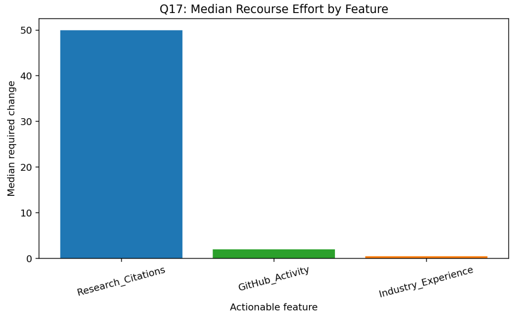

Fig. 14. Median actionable effort to flip near-boundary negative predictions.

Interpretation: Recourse burden varies by controllable feature.
Decision Impact: Guidance can prioritize low-effort, feasible interventions.
Threat: Real-world feasibility constraints may be partially unobserved.

D. Q18: Temporal Backtesting and Degradation

Mean temporal AUC: 0.5428; AUC decay: 0.0478.

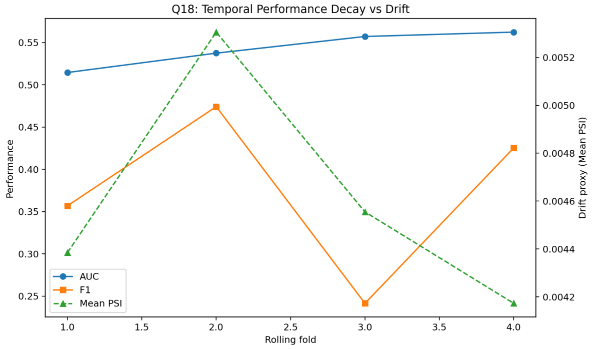

Fig. 15. Rolling temporal fold performance versus drift proxy.

Interpretation: Sequential performance degradation is measurable and drift-aware.
Decision Impact: Time-aware validation is required before future-window claims.
Threat: If true timestamps are absent, fallback ordering adds uncertainty.

E. Q19: Uncertainty Quantification

Coverage@90: 0.9000; Max under-coverage: 0.0050.

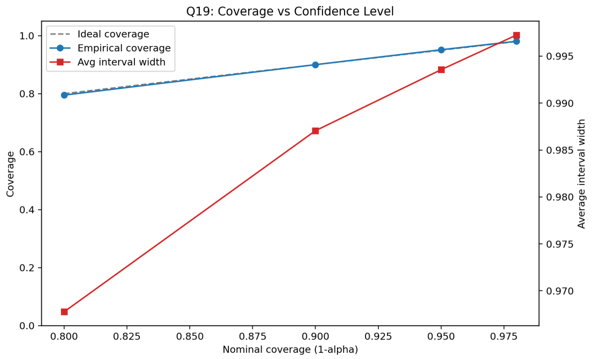

Fig. 16. Nominal vs empirical coverage and interval width.

Interpretation: Conformal coverage reflects reliability-width tradeoff.
Decision Impact: Low-confidence cases can be deferred to human review.
Threat: Distribution shift can weaken finite-sample guarantees.

F. Q20: Fairness Mitigation

DP gap (baseline): 0.1550; DP gap (mitigated): 0.0978; Policy pass: true.

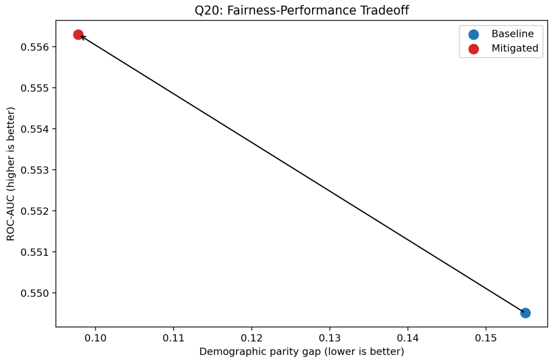

Fig. 17. Fairness–performance tradeoff after mitigation (e.g., reweighing).

Interpretation: Mitigation shifts operating point on utility-fairness plane.
Decision Impact: Deployment depends on explicit policy bounds for utility loss.
Threat: Improvement in one fairness metric may mask harms elsewhere.

V. Consolidated Metric Snapshot

TABLE I
Auto-exported key outcomes

Metric	Value
Run profile	balanced
Capstone model	XGBoost
Q6 AUC / Accuracy / F1	0.5495 / 0.5835 / 0.2475
Q15 Brier / ECE	0.2436 / 0.0327
Q15 Best-F1 threshold	0.2500
Q16 Top PSI feature (value)	Visa_Approval_Date (0.0013)
Q17 Recourse success rate	1.0000
Q18 Mean AUC / AUC decay	0.5428 / 0.0478
Q19 Coverage@90 / Max under-coverage	0.9000 / 0.0050
Q20 DP gap baseline → mitigated	0.1550 → 0.0978
Q20 Policy pass	true

VI. Reproducibility and Artifacts

The full run is profile-driven:

```
python code/scripts/full_solution_pipeline.py –
profile {fast,balanced,heavy}
```

Generated outputs include:

- figures under code/figures/ and ../figures/
- solution tables under code/solutions/
- machine-readable summaries: run_summary.json
- report-ready exports: latex_metrics.json, latex_metrics.tex
- Q18–Q20 CSV artifacts for auditability and grading

VII. Conclusion

This capstone delivers a complete, auditable workflow that connects methodological rigor with production-readiness: leakage-safe modeling, calibrated decision policy, drift surveillance, uncertainty-aware routing, and fairness-constrained mitigation. The package is suitable for graduate instruction and near-production experimentation under explicit governance controls.