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 documents an end-to-end graduate-level data science capstone on GlobalTechTalent_50k.csv. The pipeline covers data engineering, leakage diagnostics, inference, optimization, non-linear modeling, unsupervised learning, explainable AI, and production-oriented extensions (Q15–Q20). The implementation is reproducible via a profile-aware CLI and generates report-ready metrics and artifacts automatically.

I. Problem Statement and Scope

The project predicts Migration_Status for 50,000 technical professionals while enforcing methodological controls for leakage, drift, uncertainty, and fairness. The final package is organized as a full instructional+engineering deliverable: scripts, tests, notebooks, and bilingual reports.

II. Dataset Diagnostics

A. Target Balance

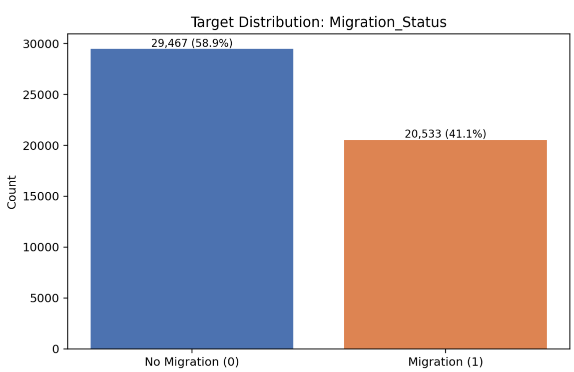

Fig. 1. Class distribution for Migration_Status.

Interpretation: The dataset is moderately imbalanced and requires threshold-aware evaluation beyond raw accuracy.
Decision Impact: Model selection and threshold policy are evaluated with AUC/F1/calibration rather than accuracy alone.
Limitation/Threat: Class prevalence can shift in deployment, so historical balance should not be treated as stationary.

B. Missingness Profile

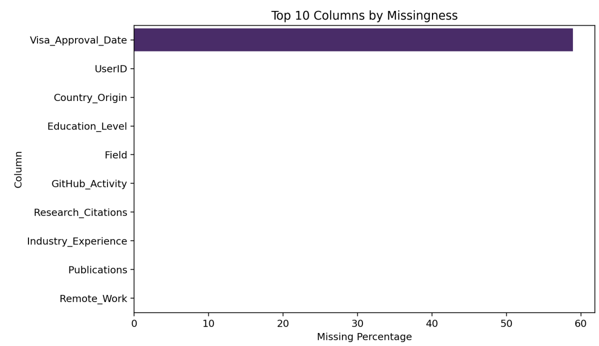

Fig. 2. Top-10 missingness rates across columns.

Interpretation: Missingness is concentrated in operational fields, notably visa-related attributes.
Decision Impact: Features tied to post-outcome process states are either removed or treated as leakage candidates.
Limitation/Threat: Missing-not-at-random behavior can encode policy effects and induce biased estimates.

C. Correlation Structure

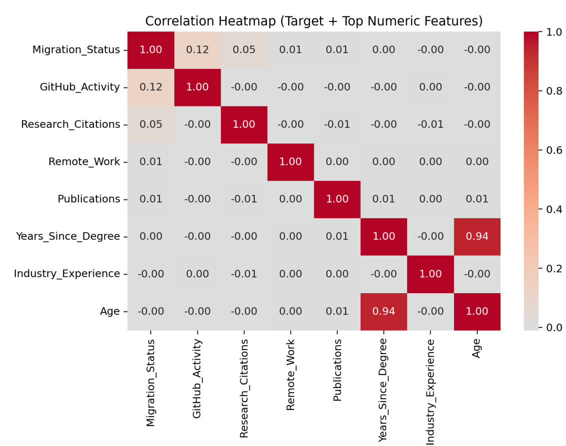

Fig. 3. Correlation heatmap for key numeric predictors and target.

Interpretation: The target correlates with activity and prominence indicators but no single predictor is sufficient.

Decision Impact: Multivariate and non-linear models are justified by mixed signal strength and interaction effects.
Limitation/Threat: Correlation is not causal; policy-sensitive confounders remain unobserved.

D. Country-Level Migration Rates

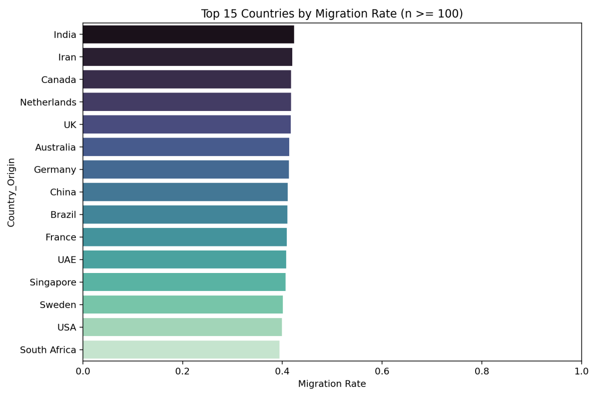

Fig. 4. Country-level migration rate ranking under minimum support filter.

Interpretation: Group-level outcome rates vary materially across origin countries.
Decision Impact: Fairness slicing and post-mitigation evaluation are mandatory before policy use.
Limitation/Threat: Observed regional differences may encode visa regimes, not individual readiness.

III. Core Questions (Q1–Q6)

A. Q1 Data Engineering and Leakage

The SQL window-function answer is exported to code/solutions/q1_moving_average.sql. Leakage diagnostics show that Visa_Approval_Date must be excluded because it is a post-outcome process artifact.

B. Q3 Optimizer Dynamics

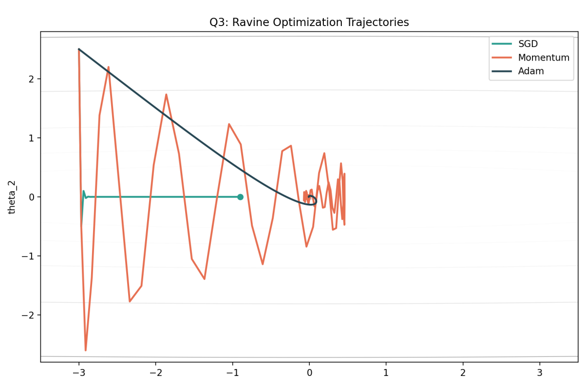

Fig. 5. SGD vs Momentum vs Adam trajectories on a ravine objective.

Interpretation: Momentum and Adam reduce ravine oscillation and converge faster than plain SGD.
Decision Impact: For ill-conditioned objectives, adaptive or momentum-based optimizers are preferred defaults.
Limitation/Threat: Toy ravine behavior does not fully capture non-convex deep objective geometry.

C. Q4 Non-Linear Models and Complexity Control

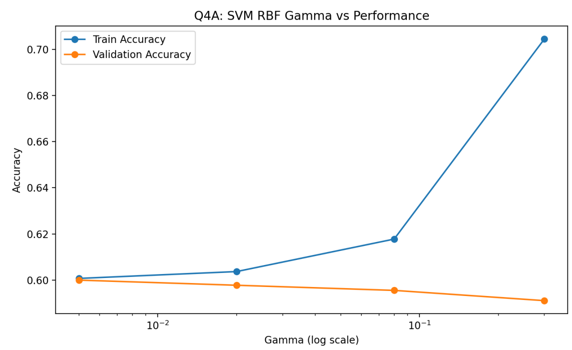

Fig. 6. Validation sensitivity to RBF kernel width (γ).

Interpretation: Larger γ increases local sensitivity and can increase variance.
Decision Impact: Overfitting regimes are controlled by reducing γ and cross-validating the support pattern.
Limitation/Threat: Hyperparameter effect depends on feature scaling and class overlap.

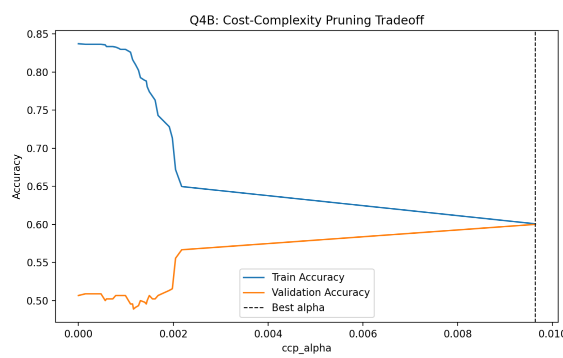

Fig. 7. Cost-complexity pruning tradeoff for CART.

Interpretation: Increasing α regularizes tree size and moves along the bias-variance frontier.
Decision Impact: Selected α should maximize validation generalization, not training fit.
Limitation/Threat: Pruning curves are data-split dependent and may drift under regime changes.

D. Q5 Unsupervised Structure

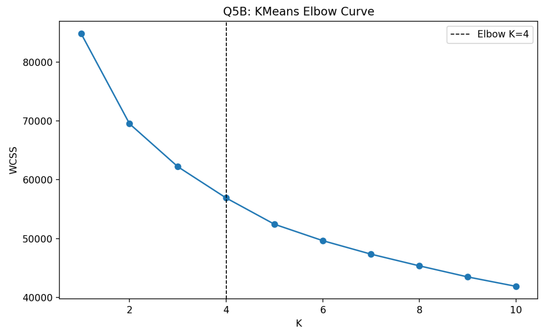

Fig. 8. WCSS elbow curve for model order selection in K-Means.

Interpretation: WCSS shows diminishing returns after a moderate cluster count.
Decision Impact: Elbow-based K is treated as a complexity heuristic, then validated against interpretability utility.
Limitation/Threat: Elbow location can be ambiguous when curvature is shallow.

E. Q6 Explainable Capstone Model

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

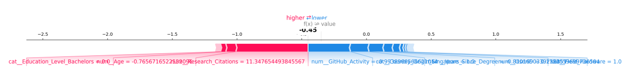

Fig. 9. Local SHAP explanation for a selected high-citation candidate.

Interpretation: The candidate prediction is the sum of feature-level positive and negative contributions from the base value.
Decision Impact: Decision review can target dominant negative drivers instead of relying on aggregate score alone.
Limitation/Threat: SHAP explains model behavior, not structural causality.

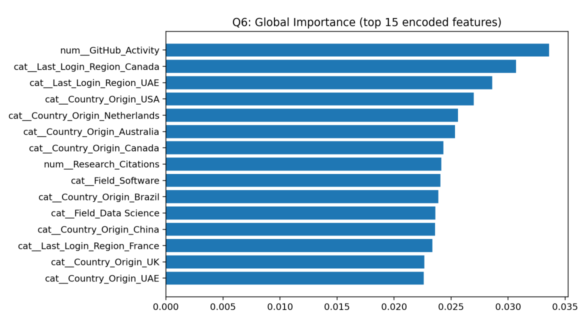

Fig. 10. Global feature influence summary for the capstone model.

Interpretation: Research and activity-related predictors dominate the global attribution profile.
Decision Impact: Feature governance should prioritize data quality and policy review for top-ranked drivers.
Limitation/Threat: Global importance does not reflect all subgroup-specific interactions.

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

A. Q15 Calibration and Threshold Policy

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

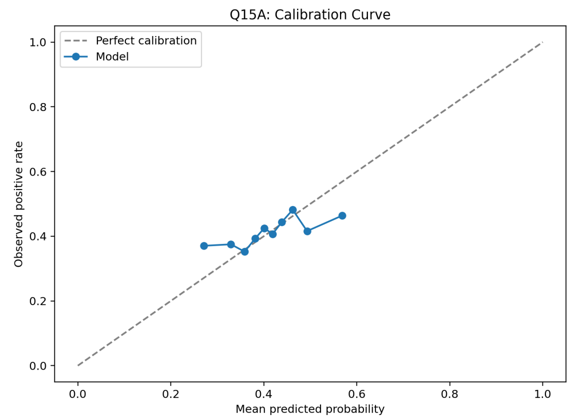

Fig. 11. Calibration reliability curve for the capstone model.

Interpretation: The calibration curve quantifies probability reliability against empirical frequencies.
Decision Impact: Operational thresholds are selected from calibrated risk estimates rather than raw scores.
Limitation/Threat: Calibration can degrade over time and must be monitored.

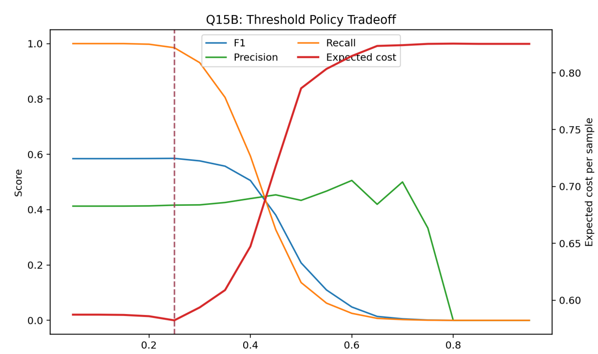

Fig. 12. Threshold tradeoff between F1 components and expected decision cost.

Interpretation: Precision, recall, and expected error cost respond differently to threshold movement.
Decision Impact: The threshold is chosen by policy utility (cost matrix), not a universal default.
Limitation/Threat: Cost assumptions are context-dependent and can change by jurisdiction.

B. Q16 Drift Monitoring

Top drift feature: Visa_Approval_Date with PSI 0.0013.

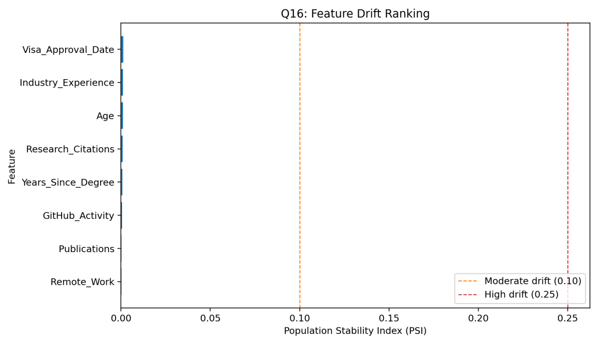

Fig. 13. Feature drift ranking using PSI thresholds.

Interpretation: A subset of features exhibits moderate-to-high instability between reference and current windows.
Decision Impact: Retraining and feature-level alerting are triggered by PSI severity bands.
Limitation/Threat: PSI detects distribution shift, not necessarily target relationship drift.

C. Q17 Counterfactual Recourse

Recourse success rate: 1.0000.

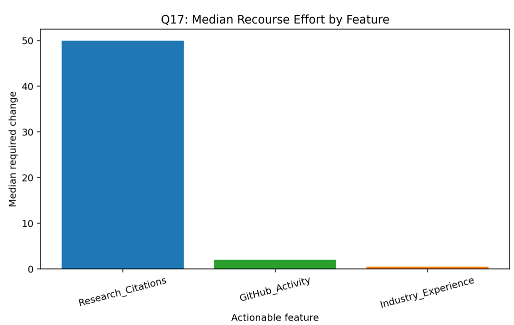

Fig. 14. Median actionable effort needed to flip near-boundary predictions.

Interpretation: Actionability varies across controllable features and can be quantified.
Decision Impact: Candidate guidance can focus on the lowest-effort feasible interventions.
Limitation/Threat: Counterfactual feasibility depends on real-world constraints not encoded in data.

D. Q18 Temporal Backtesting and Degradation

Mean temporal AUC: 0.5428, AUC decay: 0.0478.

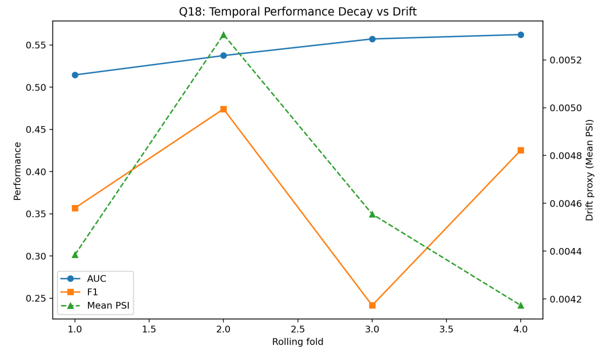

Fig. 15. Rolling temporal folds: performance decay against drift proxy.

Interpretation: Performance changes across sequential folds and is contrasted with drift magnitude.
Decision Impact: Time-aware validation is required before production claims on future windows.
Limitation/Threat: If a true time field is absent, fallback ordering introduces uncertainty in temporal realism.

E. Q19 Uncertainty Quantification

Coverage@90: 0.9000, max under-coverage: 0.0050.

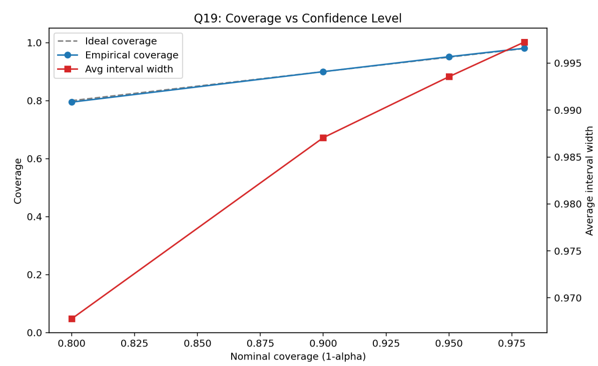

Fig. 16. Nominal vs empirical coverage and interval width across confidence levels.

Interpretation: Split-conformal intervals expose the calibration-quality tradeoff between reliability and interval width.
Decision Impact: Policy can reject low-confidence cases and route them for human review.
Limitation/Threat: Distribution shift can break finite-sample coverage guarantees.

F. Q20 Fairness Mitigation Experiment

DP gap baseline: 0.1550; DP gap mitigated: 0.0978; policy pass: true.

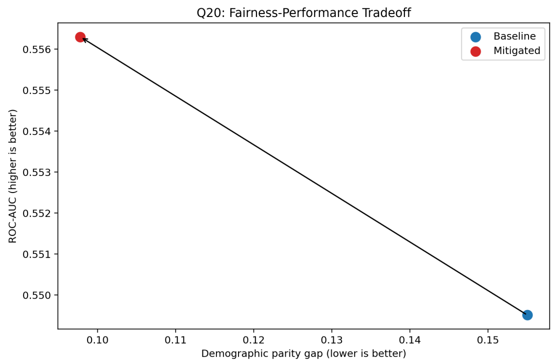

Fig. 17. Fairness-performance tradeoff from baseline to mitigated model.

Interpretation: Reweighing shifts the operating point on the fairness-performance plane.
Decision Impact: Deployment requires explicit policy constraints on acceptable performance drop and fairness gain.
Limitation/Threat: Single-metric fairness gains can mask harms on unmonitored subgroups.

V. Reproducibility and Artifacts

The implementation supports profile-based execution via python code/scripts/full_solution_pipeline.py –profile {fast,balanced,heavy}. All artifacts are generated under:
code/solutions/ and code/figures/, including schema-v2 run_summary.json, Q18–Q20 CSV outputs, and report metric exports (latex_metrics.json/.tex).

VI. Conclusion

The project now operates as a full graduate-capstone package: it combines mathematical rigor, engineering reproducibility, production diagnostics, and explainability/fairness accountability in a single auditable workflow.