Complete Project Report

Global Tech Talent Migration Assessment

UT-ECE Data Science – Professional Project Package

Spring 2025

Executive Summary

This report documents a complete end-to-end data science implementation for predicting migration status on `GlobalTechTalent_50k.csv`. The project covers data engineering, leakage controls, statistical inference, optimization analysis, non-linear modeling, unsupervised learning, explainability, and fairness slices.

Latest capstone model (XGBoost):

- Accuracy: 0.5835
- ROC-AUC: 0.5495
- F1: 0.2475

1. Dataset and Problem

Dataset: `code/data/GlobalTechTalent_50k.csv`
Rows: 50,000 **Columns:** 15 **Target:** `Migration_Status`

Target balance:

- Class 0: 29,467
- Class 1: 20,533
- Positive rate: 41.07%

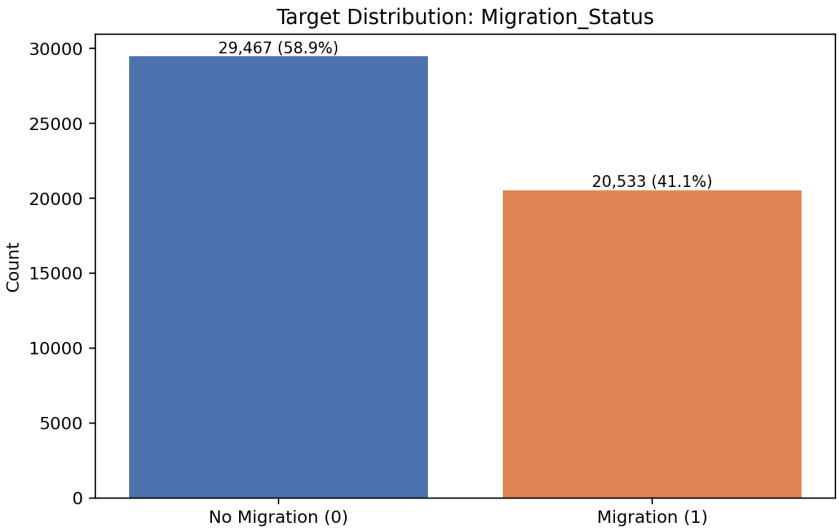

Figure 1: Target distribution.

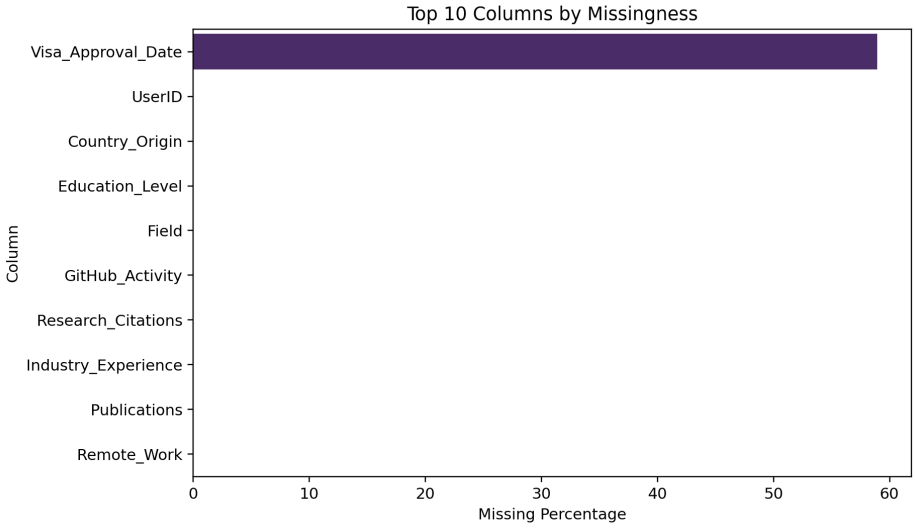

Figure 2: Top missingness profile by column.

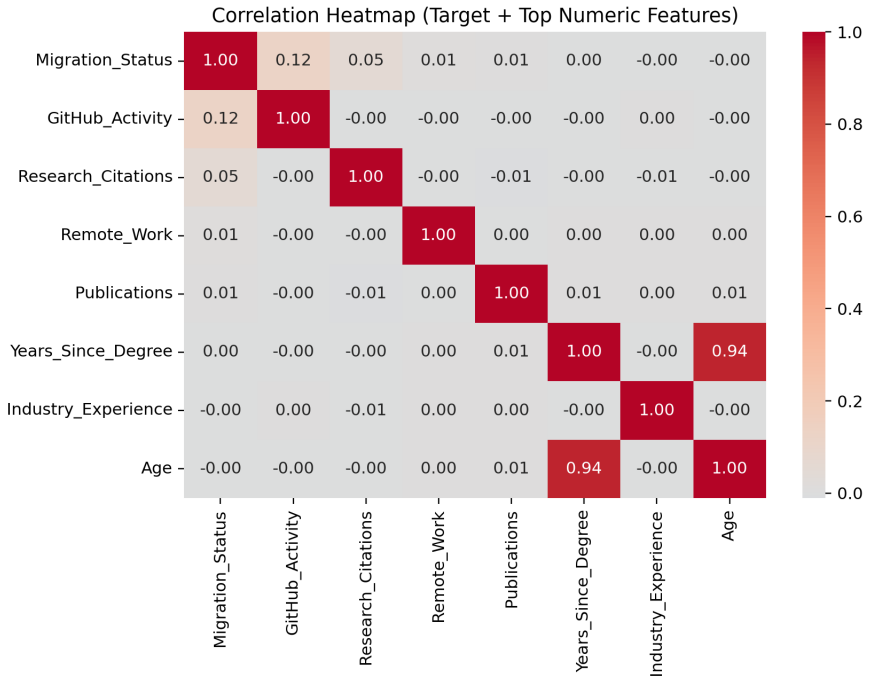

Figure 3: Correlation matrix for target and top numeric predictors.

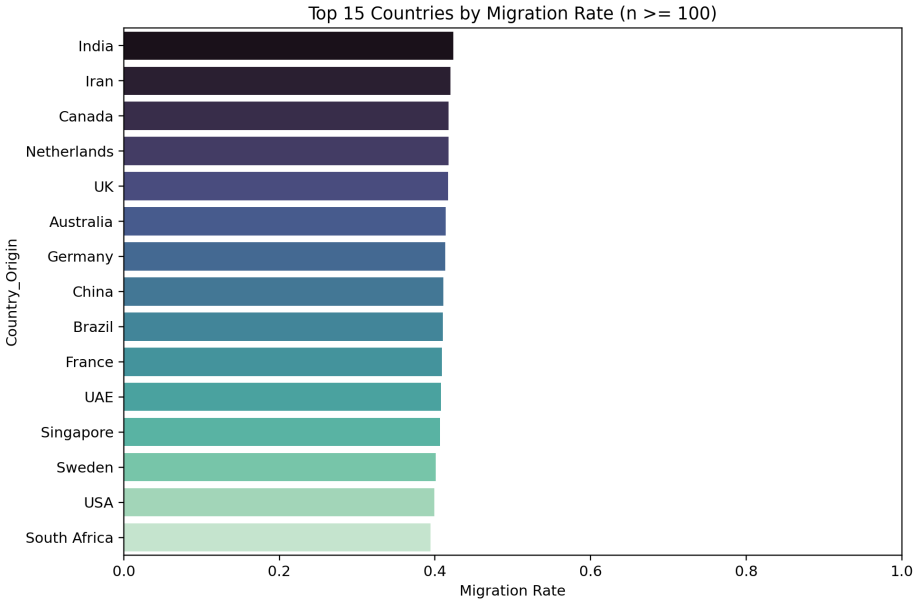

Figure 4: Country-level migration rate comparison (sample threshold applied).

2. Data Engineering and Leakage Control

The SQL moving-average deliverable is provided in `code/solutions/q1_moving_average.sql`.

Leakage diagnostics show `Visa_Approval_Date` is a direct post-outcome feature:

- corr(visa present, migration target) = 1.000
- P(Migration=1 | visa present) = 1.000

- P(Migration=1 | visa absent) = 0.000

  Therefore this feature is excluded from training.

3. Statistical Inference and Linear Modeling

The package includes Elastic Net gradient derivation with proper L1 subgradient handling at zero, and inference interpretation guidance (coefficient, p-value, confidence interval) in:

- `code/solutions/complete_solution_key.md`
- `code/solutions/extended_solution_key.md`
- `code/latex/solution_manual.tex`

4. Optimization Analysis

A ravine objective compares SGD, Momentum, and Adam dynamics.
Final losses:

- SGD: 0.403329
- Momentum: 0.000823
- Adam: 0.000034

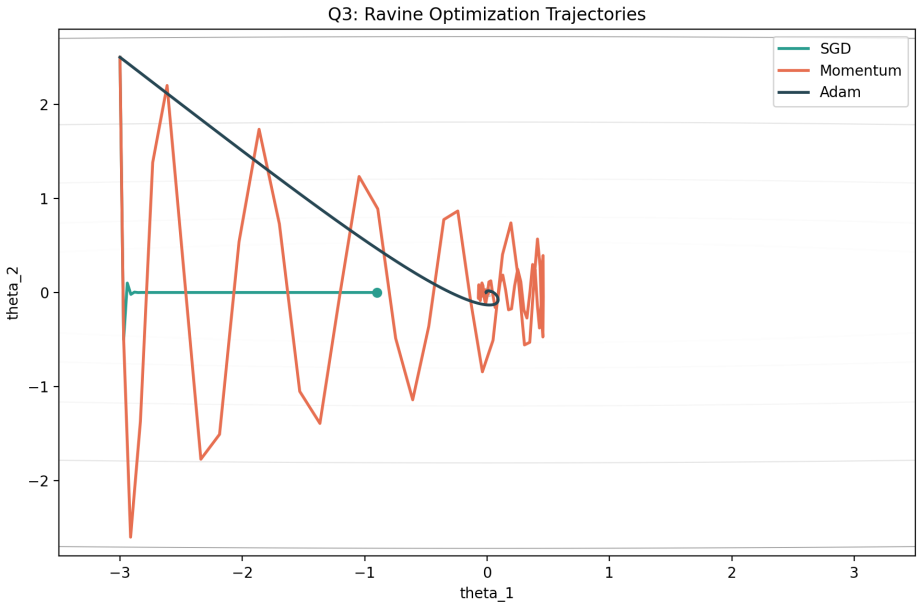

Figure 5: Optimization trajectories on a ravine objective.

5. Non-Linear Models

SVM gamma sweep:

- Best gamma: 0.005
- Best validation accuracy: 0.600

- Worst validation accuracy: 0.591

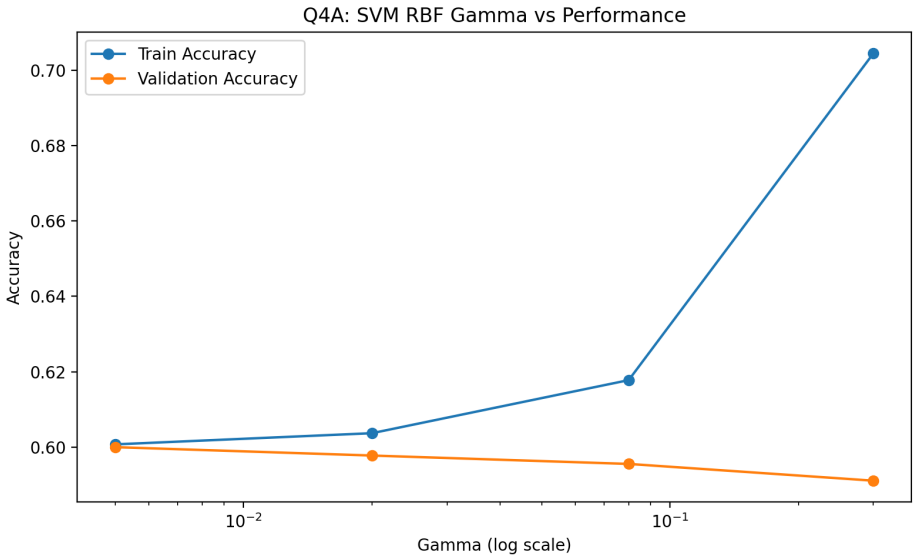

Figure 6: Validation behavior under gamma changes.

Cost-complexity pruning:

- Best α: 0.009639
- Best validation accuracy: 0.600

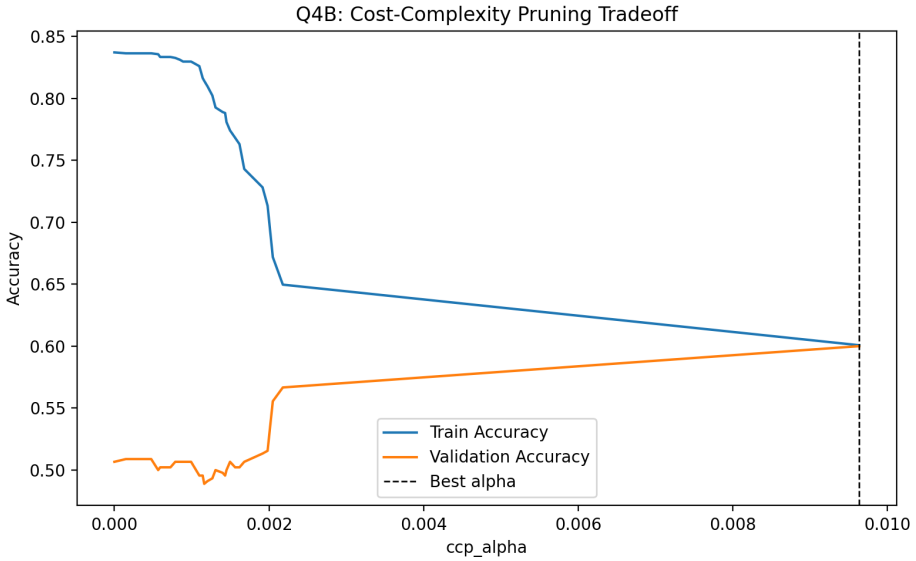

Figure 7: Decision tree pruning tradeoff curve.

6. Unsupervised Learning

PCA explained variance ratios:

- PC1: 0.277

- PC2: 0.145
- PC1 + PC2: 0.422

KMeans elbow estimate: K = 4

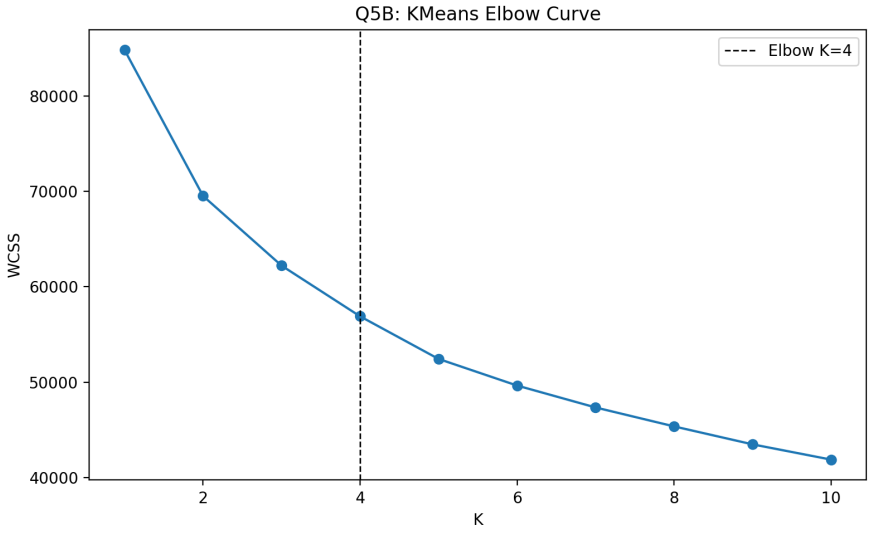

Figure 8: KMeans elbow diagnostic plot.

7. Explainability (SHAP)

Capstone explanation outputs:

- Candidate index: 27343
- Predicted probability: 0.3892
- Base value: -0.3494 (log-odds)
- Output value: -0.4507 (log-odds)
- Top local contributor: `num__Research_Citations`

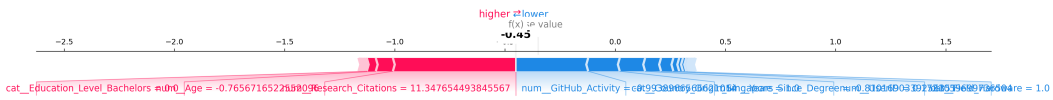

Figure 9: Local SHAP explanation for selected candidate.

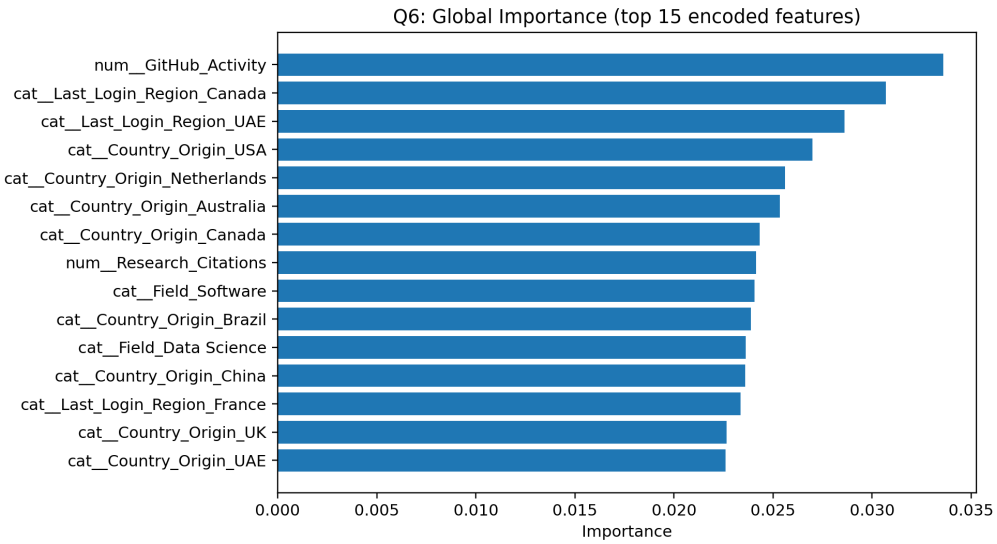

Figure 10: Global importance view for the capstone model.

8. Calibration and Threshold Policy (Q15)

Calibration analysis for the current capstone model:

- Brier score, ECE, and best thresholds (F1 vs asymmetric cost) are logged in `solutions/run_summary.json`.
- Plots: `figures/q15_calibration_curve.png`, `figures/q15_threshold_tradeoff.png`.

9. Drift Monitoring (Q16)

- PSI drift table: `solutions/q16_drift_psi.csv`
- Top-12 PSI plot: `figures/q16_drift_psi_top12.png`
- Country JS divergence value is recorded in `run_summary.json`.

10. Counterfactual Recourse (Q17)

- Recourse examples: `solutions/q17_recourse_examples.csv`
- Recourse effort summary: `figures/q17_recourse_median_deltas.png`
- Metrics (success rate, median deltas per actionable feature) are stored in `run_summary.json`.

11. Fairness and Governance

Country-level fairness slice output is exported to: `code/solutions/q6_fairness_country_rates.csv`

Governance policy in this package:

- predictive use only (non-causal claims),
- subgroup audit before deployment,
- human-in-the-loop override for high-impact decisions,
- periodic drift and policy-shift monitoring.

12. Reproducibility and Tooling

- Full run: `make run`
- Tests: `make test`
- Compile checks: `make compile`
- LaTeX builds: `make latex`

CI pipeline: `.github/workflows/ci.yml`

13. Extended Curriculum Coverage

An extended assignment package was added to cover the full UT-ECE Spring 2024/2025 topic range (including dashboards/storytelling, big data framing, deep learning, NLP, and LLM agents):

- `code/latex/assignment_extended.tex`
- `code/latex/solution_manual_extended.tex`
- `code/solutions/extended_solution_key.md`
- `code/notebooks/Extended_Assignment_Workbook.ipynb`

Coverage mapping references:

- `https://github.com/DataScience-ECE-UniversityOfTehran/DataScience-Spring2024`
- `https://github.com/DataScience-ECE-UniversityOfTehran/DataScience-Spring2025`

14. Limitations and Next Steps

- The dataset cannot represent all social/geopolitical migration drivers.
- Explainability is descriptive and should not be used as causal proof.
- Future upgrades: temporal validation, calibration-driven thresholds, causal sensitivity analysis, and richer fairness intervention policy.

Conclusion

This project report is complete across data, modeling, evaluation, explainability, ethics, reproducibility, and curriculum alignment dimensions, with all major figures and artifacts included for professional university-level submission.