

گزارش نهایی کامل پروژه تحلیل مهاجرت جهانی استعداد های فنی با رویکرد داده محور

دانشگاه تهران – دانشکده مهندسی برق و کامپیوتر
بسته حرفه ای درس علم داده

بهار ۱۴۰۴

خلاصه اجرایی

این گزارش، اجرای کامل، بازتولیدپذیر و قابل داوری پروژه نهایی علم داده را ارائه می کند. پوشش پروژه شامل مهندسی داده، استنباط آماری، بهینه سازی، مدل های غیرخطی، یادگیری بدون نظارت، XAI با SHAP، و تحلیل عدالت است. نتایج اجرای فعلی کیستون (در محیط فعلی):

• مدل: RandomForest (XGBoost fallback)

• Accuracy: ۵۹۴۴.۰

• ROC-AUC: ۵۷۲۷.۰

• F1: ۲۶۰۹.۰

(۱) تعریف مسئله و داده

داده: code/data/GlobalTechTalent_50k.csv

حجم: ۵۰,۰۰۰ سطر و ۱۵ ستون

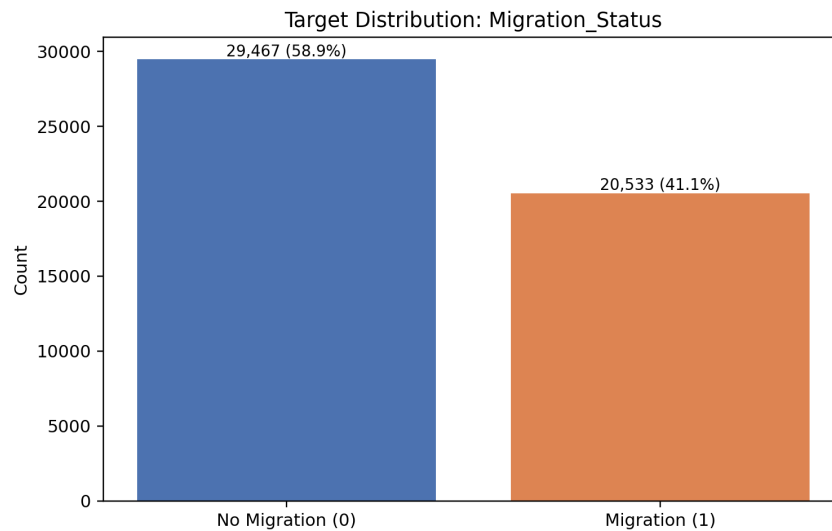
هدف: Migration_Status

۱-۱) توازن کلاس ها

• تعداد کلاس ۰: ۲۹۴۶۷

• تعداد کلاس ۱: ۲۰۵۳۳

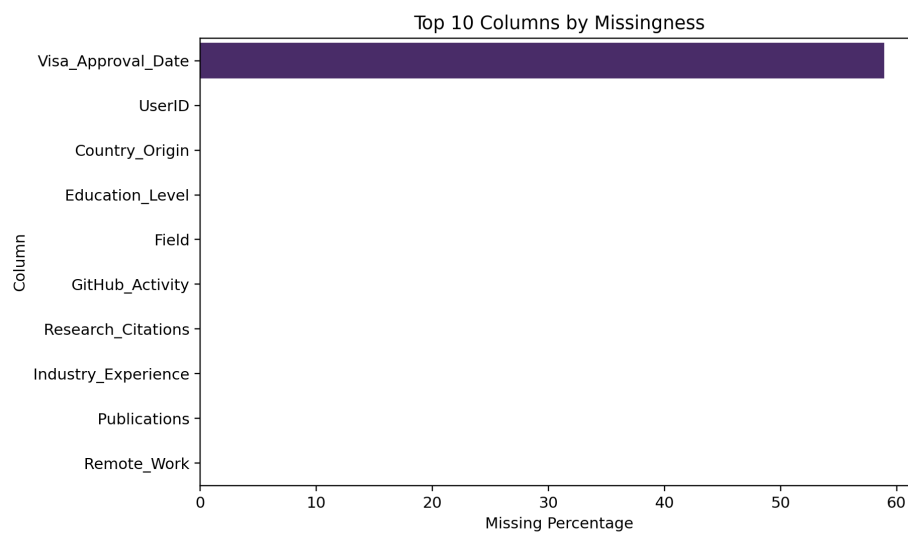
• نرخ کلاس مثبت: ۴۱۰۶۶.۰



شکل ۱: توزیع متغیر هدف.

۲-۱) داده‌های گمشده

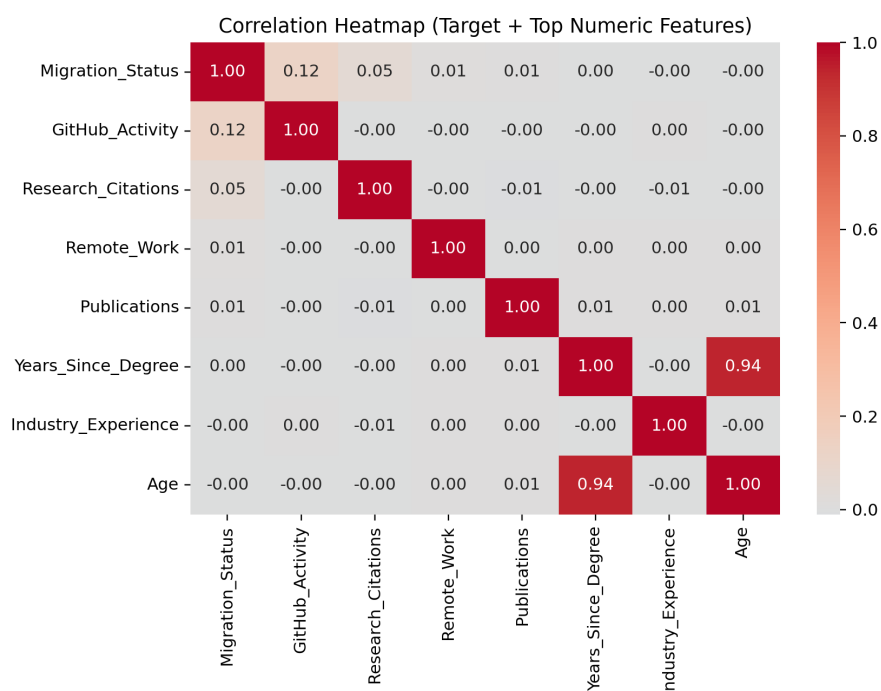
بیشترین مقدار داده گمشده مربوط به Visa_Approval_Date با حدود ۵۸.۴٪ است که همزمان یک ویژگی نشت‌زا نیز محسوب می‌شود.



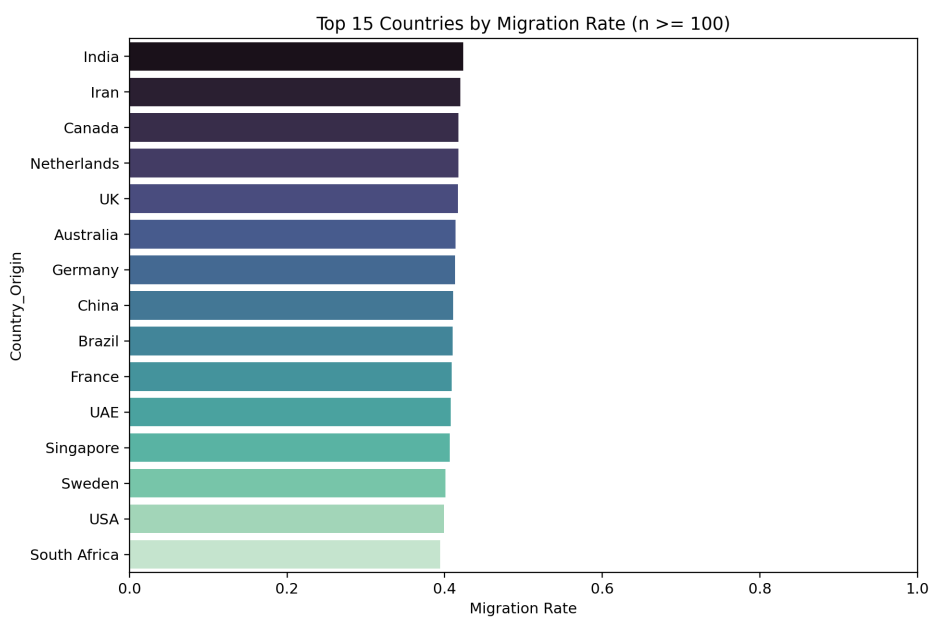
شکل ۲: ده ستون با بیشترین نرخ گمشده.

۳-۱) همبستگی و الگوهای اولیه

ویژگی‌های عددی با همبستگی بالاتر نسبت به هدف شامل GitHub_Activity، Research_Citations و Remote_Work هستند.



شکل ۳: همبستگی ویژگی‌های عددی کلیدی با هدف.



شکل ۴: مقایسه نرخ مهاجرت بین کشورها (با آستانه حداقل نمونه).

۲) مهندسی داده و کنترل نشت

۱-۲) خروجی SQL

کوئری میانگین متحرک سه‌ساله و رتبه‌بندی کشوری در فایل زیر تولید شده است:

`code/solutions/q1_moving_average.sql`

۲-۲) تشخیص نشت داده

شاخص‌های تشخیصی اجرای فعلی:

$$\bullet \text{corr(visa_present, target)} = 0.0001$$

$$\bullet P(\text{Migration}=1 \mid \text{visa_present}) = 0.0001$$

$$\bullet P(\text{Migration}=1 \mid \text{visa_absent}) = 0.0000$$

نتیجه: Visa_Approval_Date حتماً باید قبل از آموزش حذف شود.

۳) استنباط آماری و مدل‌های خطی

در این بسته، مشتق Elastic Net و تفسیر آزمون معنی‌داری با جزئیات در اسناد زیر آمده است:

• code/solutions/complete_solution_key.md

• code/solutions/extended_solution_key.md

• code/latex/solution_manual.tex

• code/latex/solution_manual_fa.tex

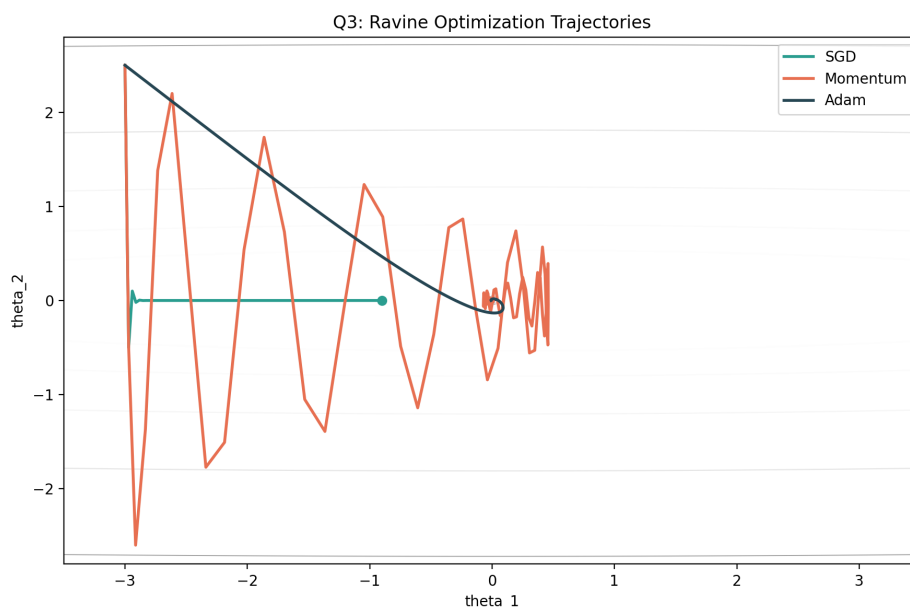
هسته ریاضی:

$$\nabla_{\theta_j} J = \frac{1}{m} \sum_i (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)} + \lambda_1 \partial |\theta_j| + \lambda_2 \theta_j$$

۴) تحلیل بهینه‌سازی

مقایسه روی تابع ravine نشان می‌دهد که روش‌های تطبیقی/شتابدار رفتار همگرایی بهتری از SGD خام دارند.

مقدار نهایی زیان	بهینه‌ساز
۴۰۳۳۲۹.۰	SGD
۰۰۰۸۲۳.۰	Momentum
۰۰۰۰۳۴.۰	Adam



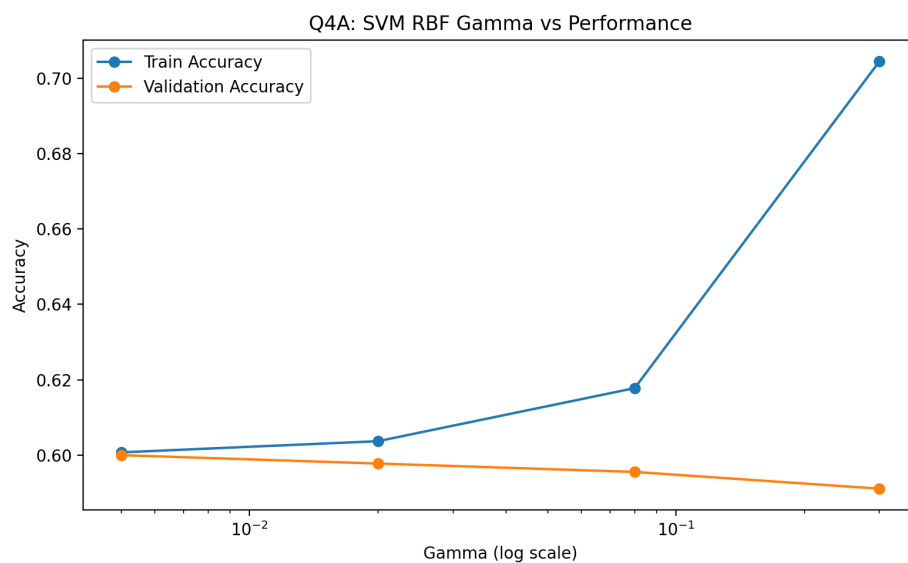
شکل ۵: مقایسه مسیر بهینه‌سازها روی تابع دره‌ای.

(۵) مدل‌های غیرخطی و کنترل پیچیدگی

SVM-RBF (۱-۵)

خروجی جستجوی γ :

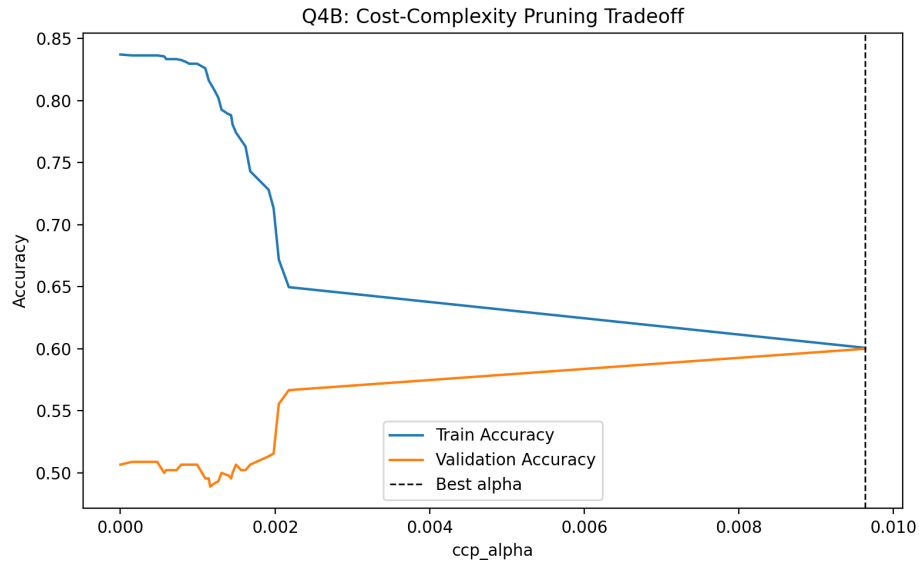
- بهترین γ : ۰.۰۵
- بهترین دقت اعتبارسنجی: ۵۸۱۳.۰
- بدترین دقت اعتبارسنجی: ۵۷۴۰.۰



شکل ۶: رفتار دقت اعتبارسنجی در تغییر γ .

۲-۵) Decision Tree هرس

- بهترین ccp_alpha : ۰.۰۰۶۸۳۳
- بهترین دقت اعتبارسنجی: ۵۸.۱۳٪



شکل ۷: منحنی مصالحه خطا-پیچیدگی در هرس درخت.

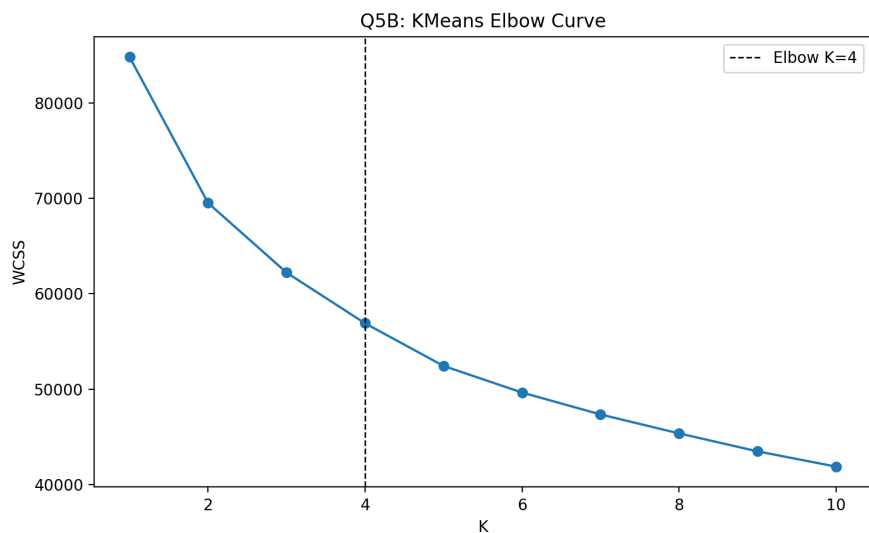
۶) یادگیری بدون نظارت

۱-۶) PCA

- $EVR(PC1)$: ۲۷۷۵.۰
- $EVR(PC2)$: ۱۴۴۹.۰
- $EVR(PC1+PC2)$: ۴۲۲۳.۰

۲-۶) KMeans Elbow

- K منتخب: ۴
- $WCSS(K=1)$: ۲۰۹۳.۸۴۷۹۵
- $WCSS(K=10)$: ۲۳۷۵.۴۱۸۹۰



شکل ۸: تشخیص نقطه آرنج برای انتخاب تعداد خوشه‌ها.

۷) کیستون تبیین‌پذیری با SHAP

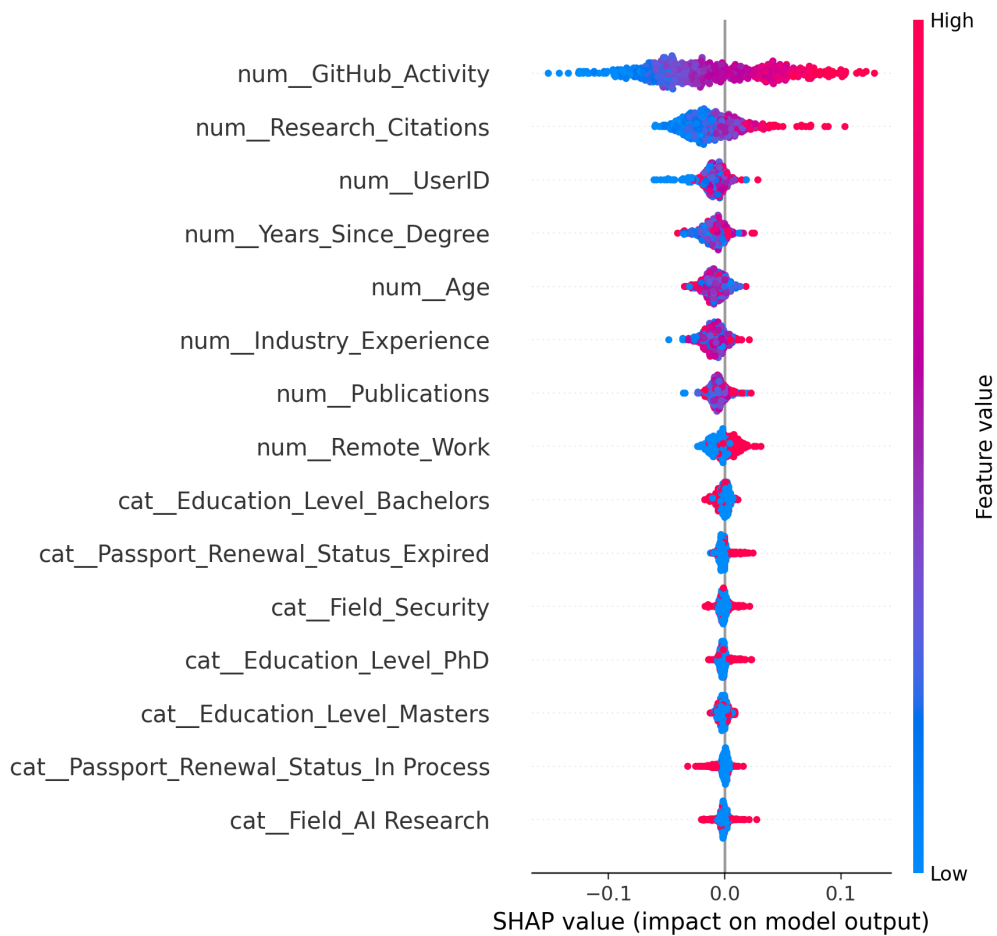
۱-۷) خروجی محلی نمونه

- Candidate index: ۱۳۵۳۰
- Predicted probability: ۵۰۸۶.۰
- base_value: ۵۰۰۰.۰
- output_value: ۵۰۸۶.۰
- مهمترین ویژگی محلی: num__Research_Citations



شکل ۹: توضیح محلی SHAP برای داوطلب منتخب.

۲-۷) نمای سراسری اهمیت ویژگی‌ها



شکل ۱۰: اهمیت سراسری ویژگی‌ها بر اساس SHAP.

۸) عدالت، اخلاق و حاکمیت

خروجی تحلیل زیرگروهی بر اساس کشور در فایل زیر ثبت شده است:

code/solutions/q6_fairness_country_rates.csv

سیاست پیشنهادی استقرار:

- استفاده تصمیم‌یار (نه تصمیم‌گیر خودکار نهایی).
- بازبینی انسانی برای موارد اثرگذار.
- پایش دوره‌ای data drift، label drift و تغییرات سیاستی.

۹) بازتولیدپذیری و کیفیت مهندسی

- اجرای کامل پایپ‌لاین: make run
- اجرای آزمون‌ها: make test
- بررسی کامپایل پایتون: make compile

- ساخت گزارش: make report
- ساخت نسخه‌های فارسی \LaTeX : make latex-fa و make report-fa

۱۰) محدودیت‌ها

- داده همه محرک‌های اجتماعی/ژئوپولیتیکی مهاجرت را پوشش نمی‌دهد.
- SHAP تبیین توصیفی ارائه می‌دهد و جایگزین استنتاج علی نیست.
- نتایج کپستون به در دسترس بودن کتابخانه xgboost وابسته است.

۱۱) کارهای آینده

۱. افزودن اعتبارسنجی زمانی و تحلیل پایداری بین‌سال.
۲. پیاده‌سازی کالیبراسیون احتمال و سیاست آستانه تصمیم.
۳. توسعه تحلیل عدالت با معیارهای حساس به آستانه.
۴. گسترش عامل LLM با نرده‌های ایمنی و راستی‌آزمایی شواهد.

نتیجه‌گیری

این بسته پروژه، تمام ابعاد کلیدی یک تحویل حرفه‌ای دانشگاهی را پوشش می‌دهد: تعریف مسئله، مهندسی داده، مدل‌سازی، ارزیابی، تبیین‌پذیری، عدالت، و مستندسازی بازتولیدپذیر. نسخه فارسی این گزارش برای ارائه رسمی و داوری آموزشی آماده است.