

گزارش نهایی کامل پروژه تحلیل مهاجرت جهانی استعدادهای فنی با رویکرد داده‌محور

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

۱۴۰۴ بهار

خلاصه اجرایی

این گزارش، اجرای کامل، بازتولیدپذیر و قابل داوری پروژه نهایی علم داده را ارائه می‌کند. پوشش پروژه شامل مهندسی داده، استنباط آماری، بهینه‌سازی، مدل‌های غیرخطی، یادگیری بدون نظارت، SHAP با XAI، و تحلیل عدالت است.

نتایج اجرای فعلی کپسون (در محیط فعلی):

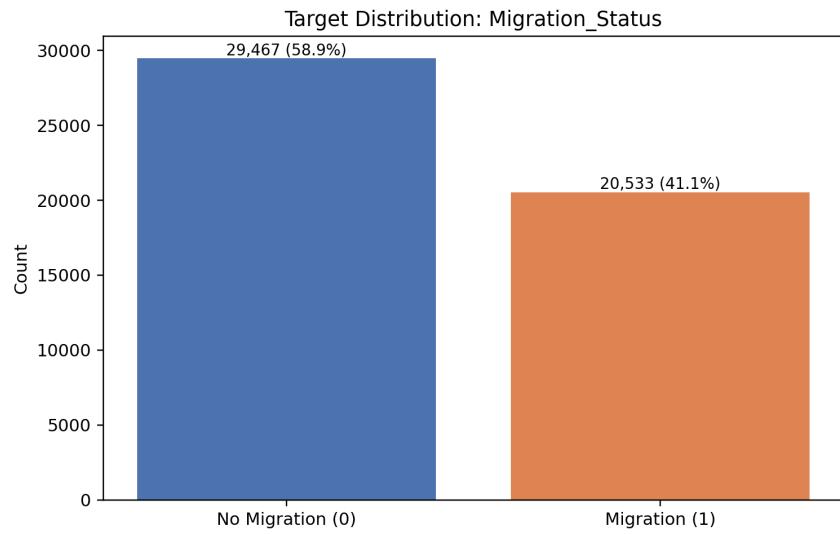
- مدل: RandomForest (XGBoost fallback)
- Accuracy: ۵۹۴۴٪
- ROC-AUC: ۵۷۲۷٪
- F1: ۲۶۰۹٪

۱) تعریف مسئله و داده

داده: code/data/GlobalTechTalent_50k.csv
حجم: ۵۰,۰۰۰ سطر و ۱۵ ستون
هدف: Migration_Status

۱-۱) توازن کلاس‌ها

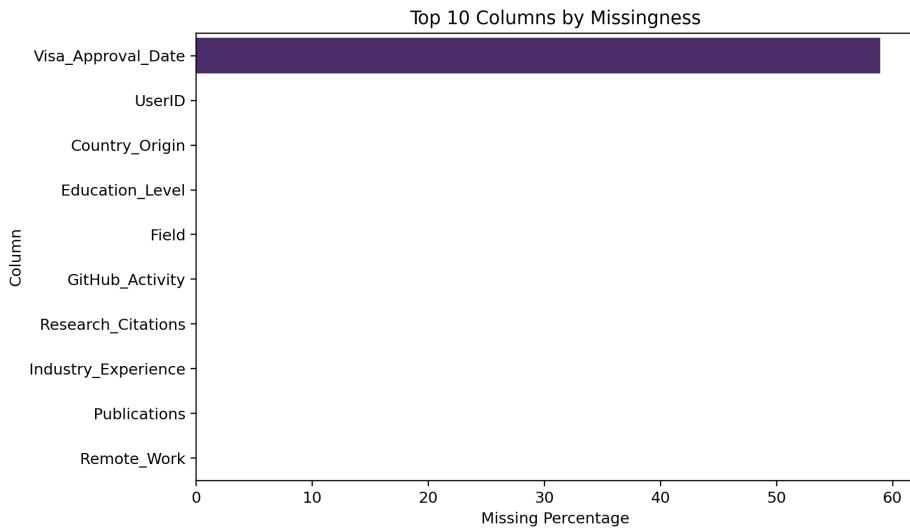
- تعداد کلاس ۰: ۲۹۴۶۷
- تعداد کلاس ۱: ۲۰۵۳۳
- نرخ کلاس مثبت: ۴۱۰۶۶٪



شکل ۱: توزیع متغیر هدف.

۲-۱) داده‌های گمشده

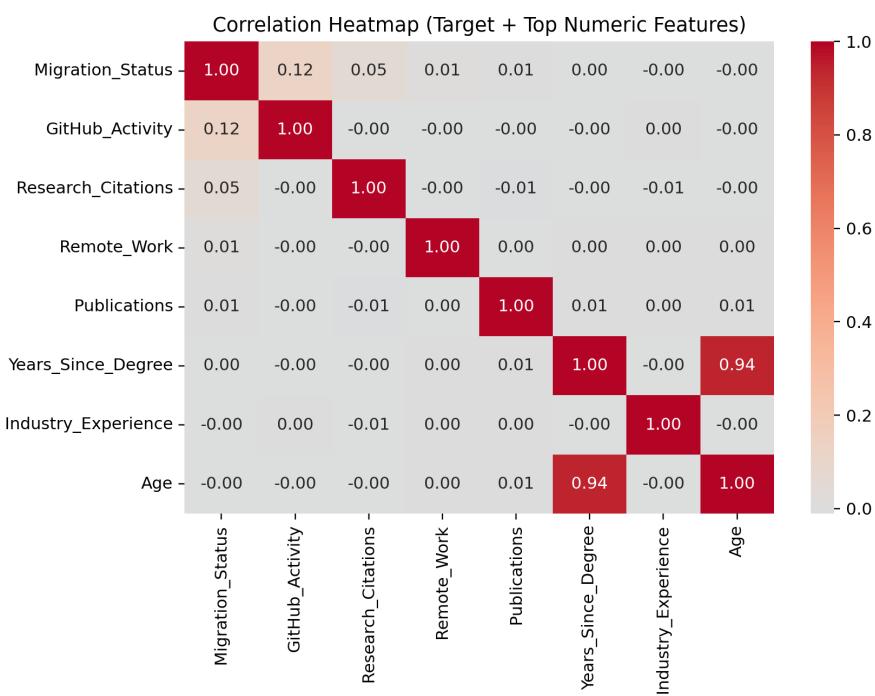
بیشترین مقدار داده گمشده مربوط به Visa_Approval_Date است که همزمان یک ویژگی نشستزانیز محسوب می‌شود.



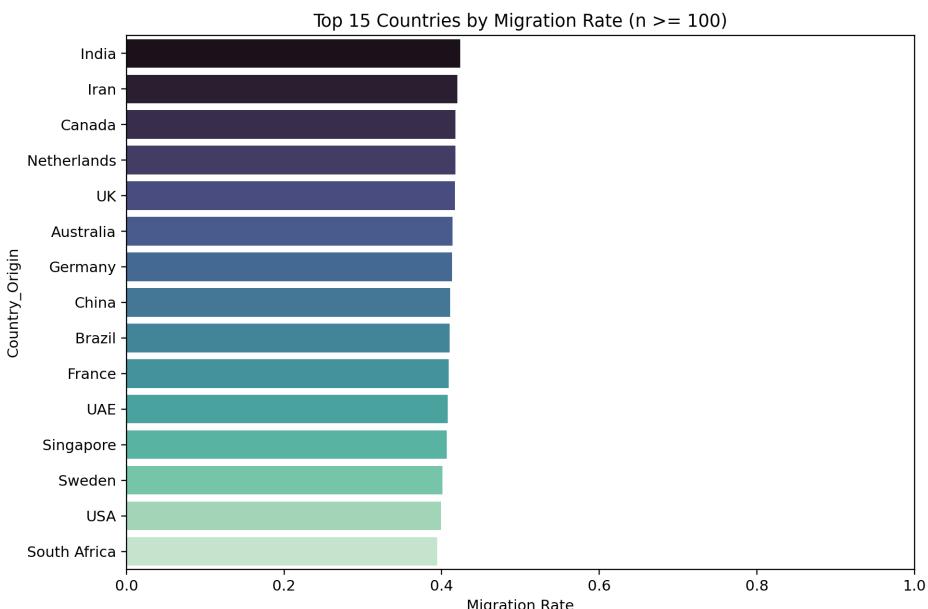
شکل ۲: ده ستون با بیشترین نرخ گمشده.

۳-۱) همبستگی و الگوهای اولیه

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



شکل ۳: همپستگی و پیزگی‌های 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 سه‌ماهه و رتبه‌بندی کشوری در فایل زیر تولید شده است:

code/solutions/q1_moving_average.sql

۲-۲) تشخیص نشت داده

شاخص‌های تشخیصی اجرای فعلی:

$$\dots_1 = \text{corr}(\text{visa_present}, \text{target}) \quad \bullet$$

$$\dots_1 = P(\text{Migration}=1 | \text{visa_present}) \quad \bullet$$

$$\dots_0 = P(\text{Migration}=1 | \text{visa_absent}) \quad \bullet$$

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:

$$\text{code/solutions/complete_solution_key.md} \quad \bullet$$

$$\text{code/solutions/extended_solution_key.md} \quad \bullet$$

$$\text{code/latex/solution_manual.tex} \quad \bullet$$

$$\text{code/latex/solution_manual_fa.tex} \quad \bullet$$

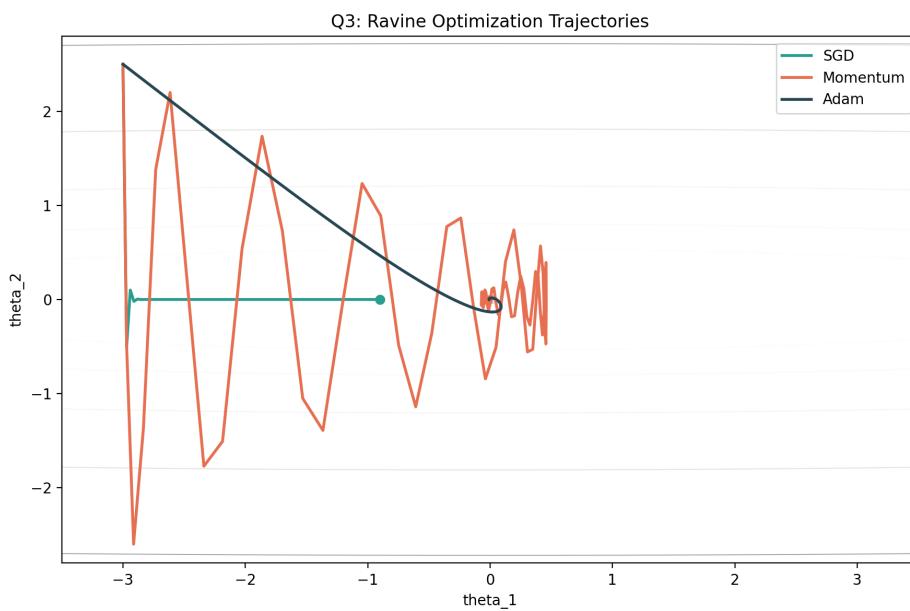
هسته ریاضی:

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

۴) تحلیل بهینه‌سازی

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

بهینه‌ساز	مقدار نهایی زیان
SGD	۴۰۳۳۲۹.۰
Momentum	۰۰۰۸۲۳.۰
Adam	۰۰۰۳۴.۰



شکل ۵: مقایسه مسیر بهینه‌سازها روی تابع دره‌ای.

۵) مدل‌های غیرخطی و کنترل پیچیدگی

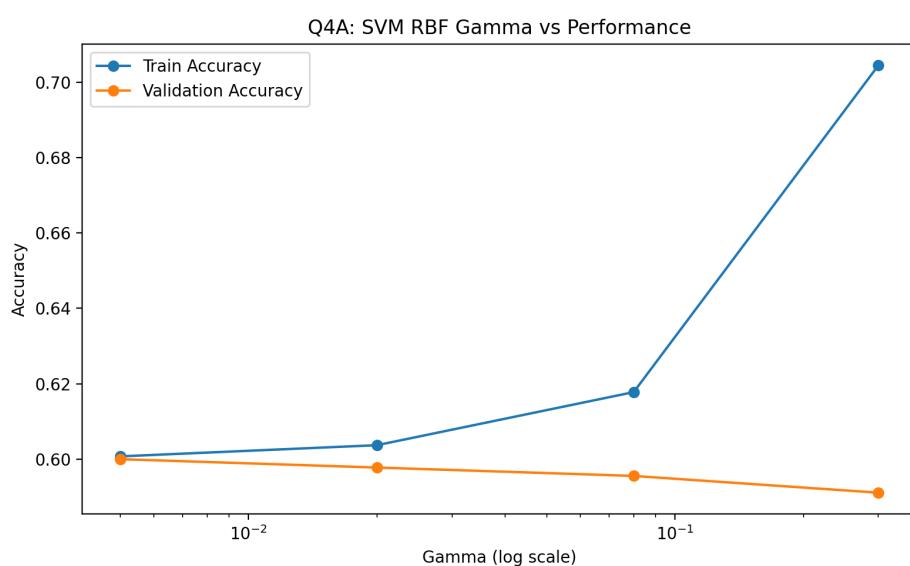
SVM-RBF (۱-۵)

خروجی جستجوی γ :

- بهترین γ : ۰۰۵۰

- بهترین دقت اعتبارسنجی: ۵۸۱۳۰

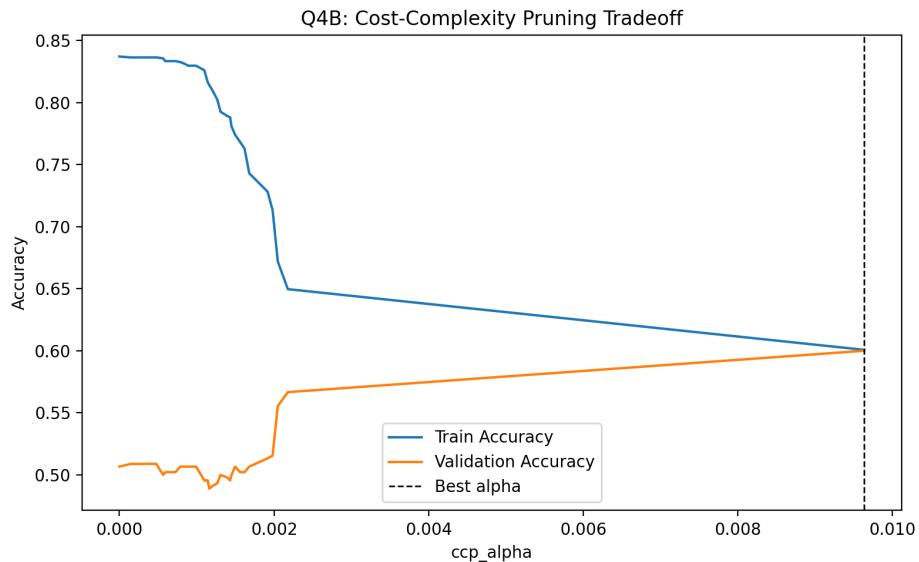
- بدترین دقت اعتبارسنجی: ۵۷۴۰۰



شکل ۶: رفتار دقت اعتبارسنجی در تغییر γ .

Decision Tree (۴-۵)

- بهترین ccp_alpha : ۰.۶۸۳۳
- بهترین دقت اعتبارسنجی: ۵۸۱۳



شکل ۷: منحنی مصالحه خطای پیچیدگی در هرس درخت.

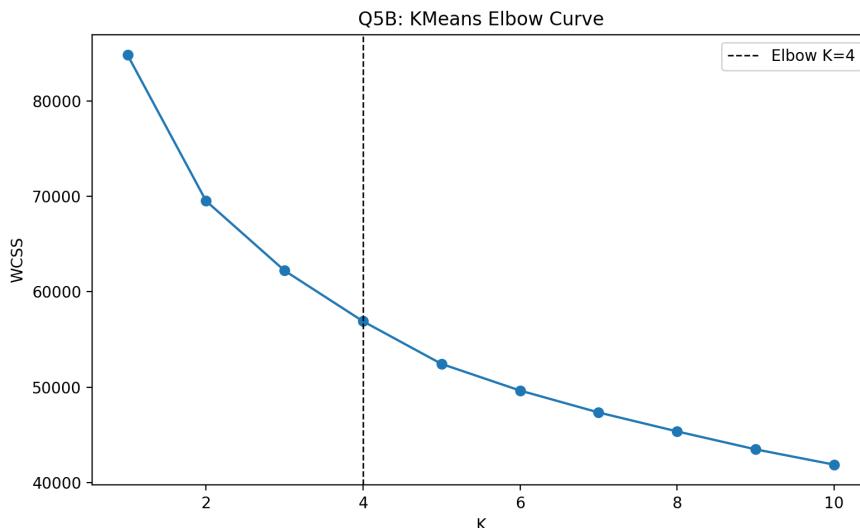
۶) یادگیری بدون نظارت

PCA (۱-۶)

- EVR(PC1): ۲۷۷۵
- EVR(PC2): ۱۴۴۹
- EVR(PC1+PC2): ۴۲۲۳

KMeans Elbow (۴-۶)

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



شکل ۸: تشخیص نقطه آرنج برای انتخاب تعداد خوشها.

۷) کپستون تبیین‌پذیری با SHAP

۱-۷) خروجی محلی نمونه

۱۳۵۳۰ :Candidate index •

5.19. Predicted probability •

5..... :base_value •

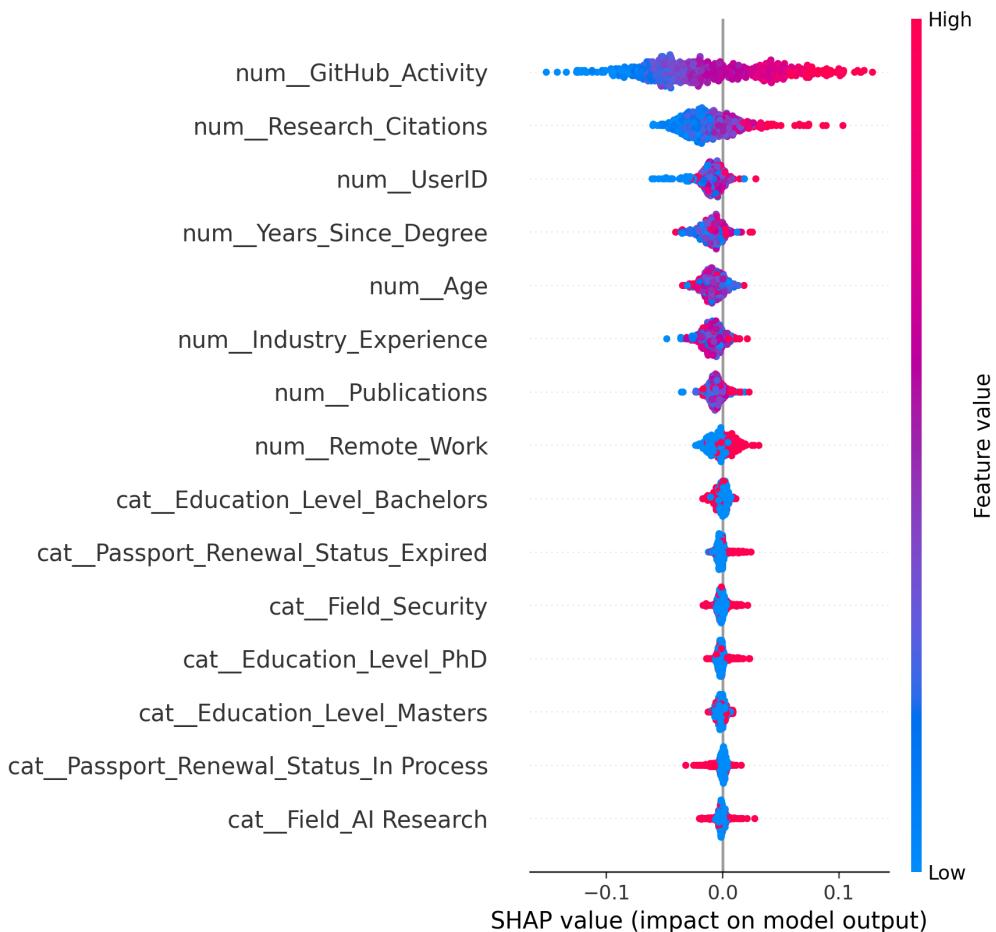
5.19. • :output_value •

• مہمترین ویژگی محلی: num_Research_Citations



شکل ۹: توضیح محلی SHAP برای داوطلب منتخب.

۷-۲) 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 تصمیمگیر (نه تصمیمگیر خودکار نهایی).
- بازبینی انسانی برای موارد اثربخش.
- پایش دوره‌ای label drift، data 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 report-fa و make latex-fa

(۱۰) محدودیت‌ها

- داده همه محرك‌های اجتماعی/ژنوپولیتیکی مهاجرت را پوشش نمی‌دهد.
- SHAP تبیین توصیفی ارائه می‌دهد و جایگزین استنتاج علی نیست.
- نتایج کپیتون 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 اینمنی و راستی‌آزمایی شواهد.

نتیجه‌گیری

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