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

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

بهار ۱۴۰۴

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، XAI با SHAP، عدالت، و سه بخش پایش تولید (کالیبراسیون، درفت، ریکورس) است.

نتایج آخرین اجرای کپستون (از run_summary.json):

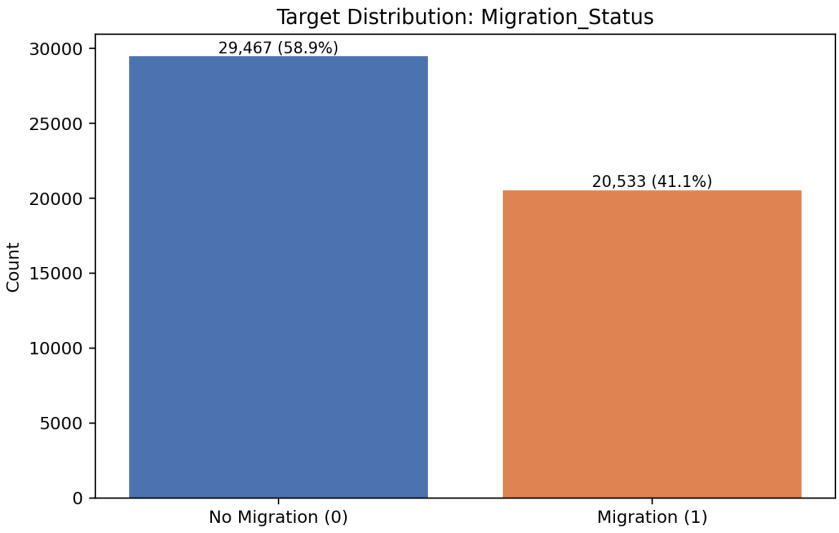
- مدل: RandomForest (XGBoost fallback)
- Accuracy: ۰.۵۹۴۴
- ROC-AUC: ۰.۵۷۲۷
- F1: ۰.۲۶۰۹

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

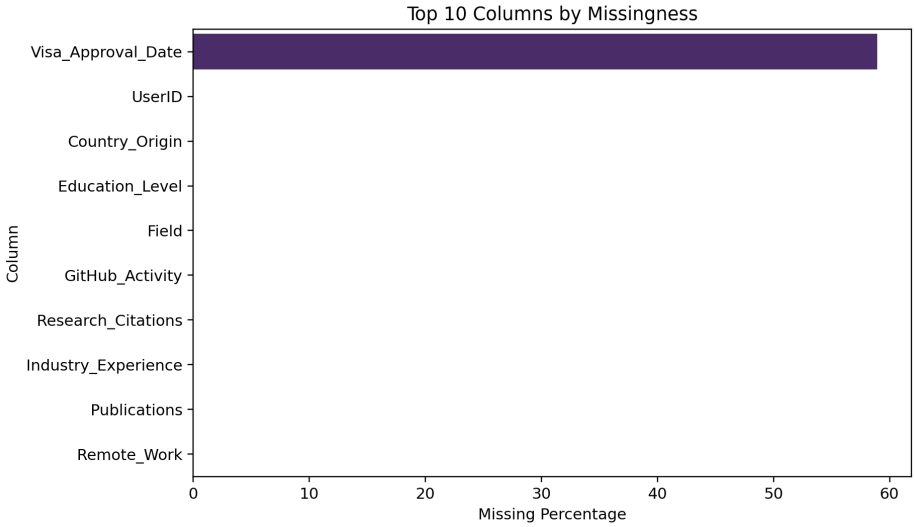
- تعداد کلاس ۰: ۲۹۴۶۷
- تعداد کلاس ۱: ۲۰۵۳۳
- نرخ کلاس مثبت: ۰.۴۱۰۶۶



شکل ۱: توزیع متغیر هدف.

۱-۲) داده‌های گمشده

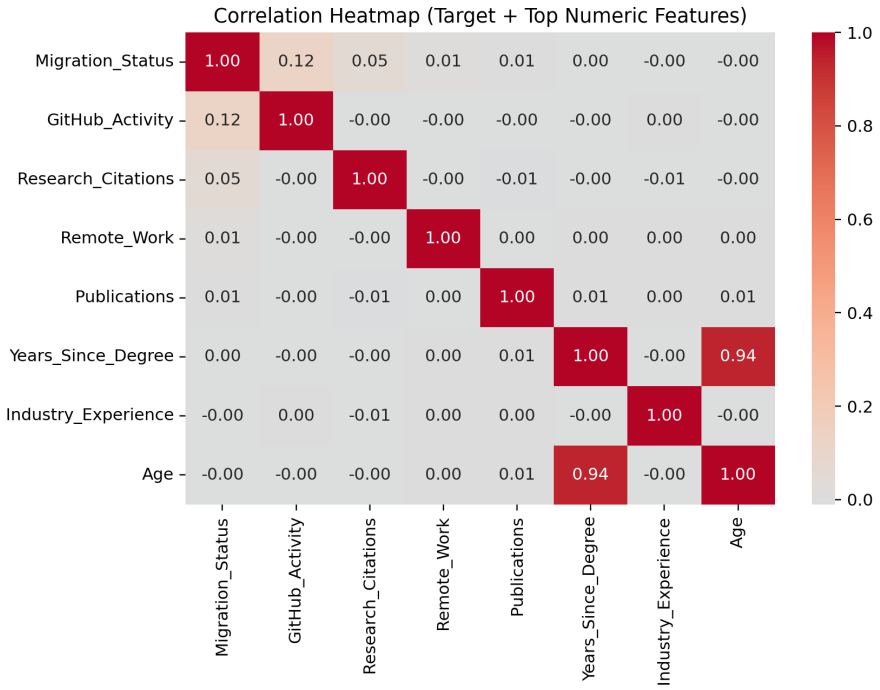
بیشترین مقدار داده گمشده مربوط به Visa_Approval_Date با حدود ۹۳۴%.۵۸ است که هم‌زمان یک ویژگی نشت‌زا نیز محسوب می‌شود.



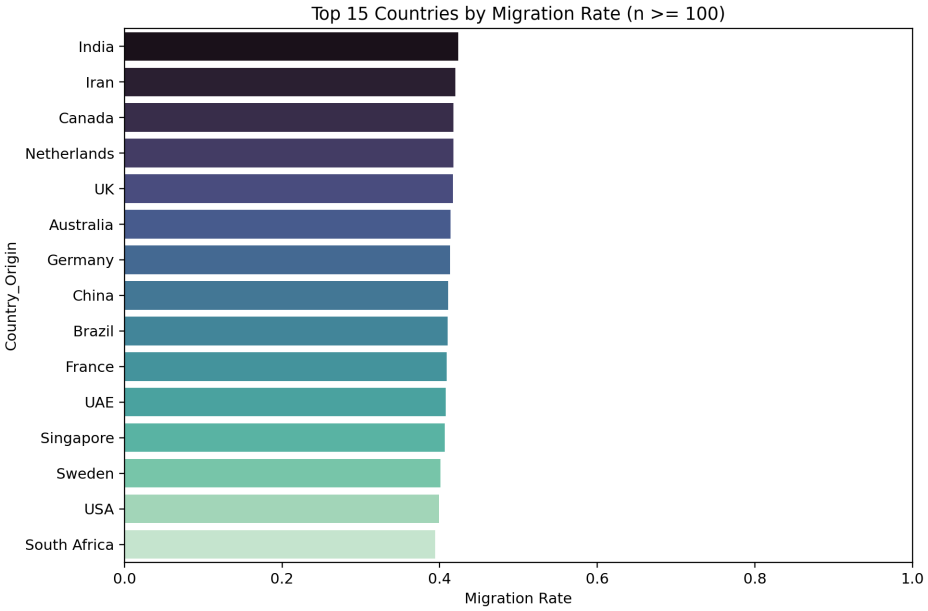
شکل ۲: ده ستون با بیشترین نرخ گمشده.

۱-۳) 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

۲-۱) 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:

- corr(visa_present, target) = ۰.۰۰۰۱
- P(Migration=1 | visa_present) = ۰.۰۰۰۱
- P(Migration=1 | visa_absent) = ۰.۰۰۰۰

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

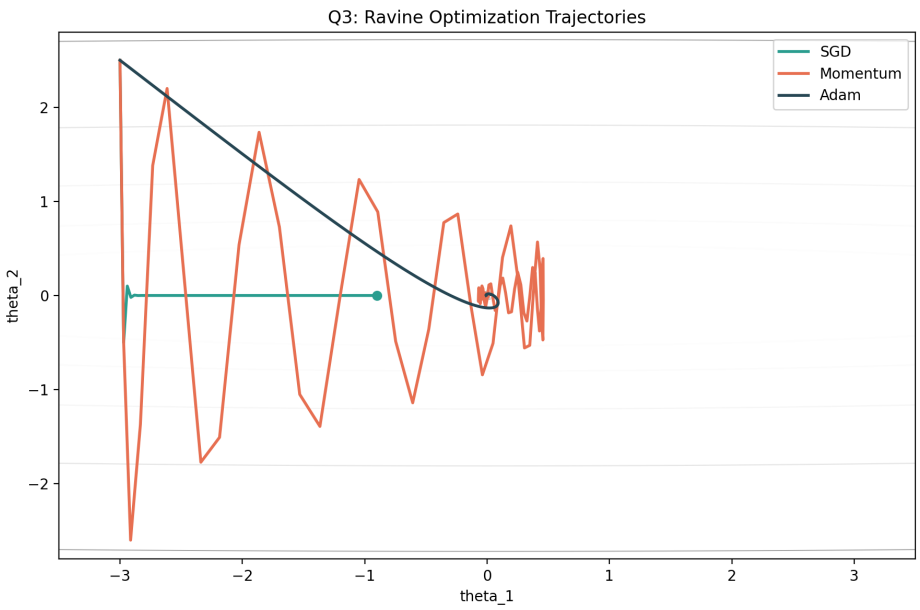
هسته ریاضی:

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

۴) تحلیل بهینه‌سازی

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

بهینه‌ساز	مقدار نهایی زیان
SGD	۴۰۳۳۲۹.۰
Momentum	۰۰۰۸۲۳.۰
Adam	۰۰۰۰۳۴.۰



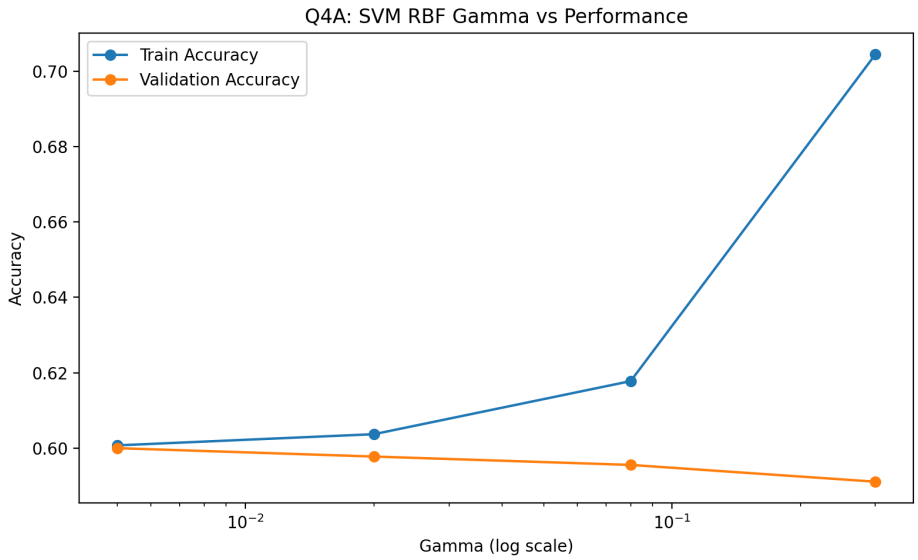
شکل ۵: مقایسه مسیر بهینه‌سازها روی تابع دره‌ای.

۵) مدل‌های غیرخطی و کنترل پیچیدگی

۵-۱) SVM-RBF

خروجی جست‌وجوی γ:

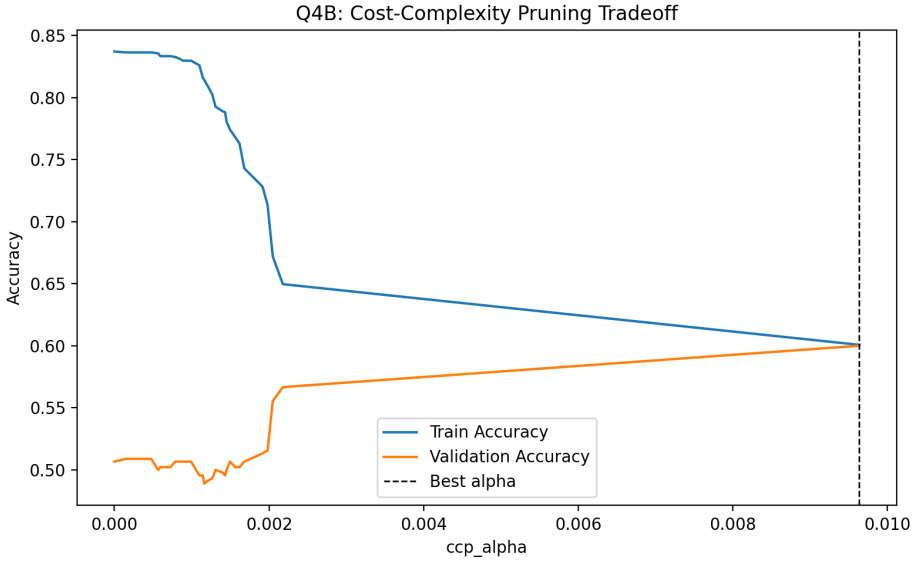
- بهترین γ: ۰.۰۰۵
- بهترین دقت اعتبارسنجی: ۰.۵۸۱۳
- بدترین دقت اعتبارسنجی: ۰.۵۷۴۰



شکل ۶: رفتار دقت اعتبارسنجی در تغییر γ.

۵-۲) هرس Decision Tree

- بهترین ccp_alpha: ۰.۰۰۶۸۳۳
- بهترین دقت اعتبارسنجی: ۰.۵۸۱۳



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

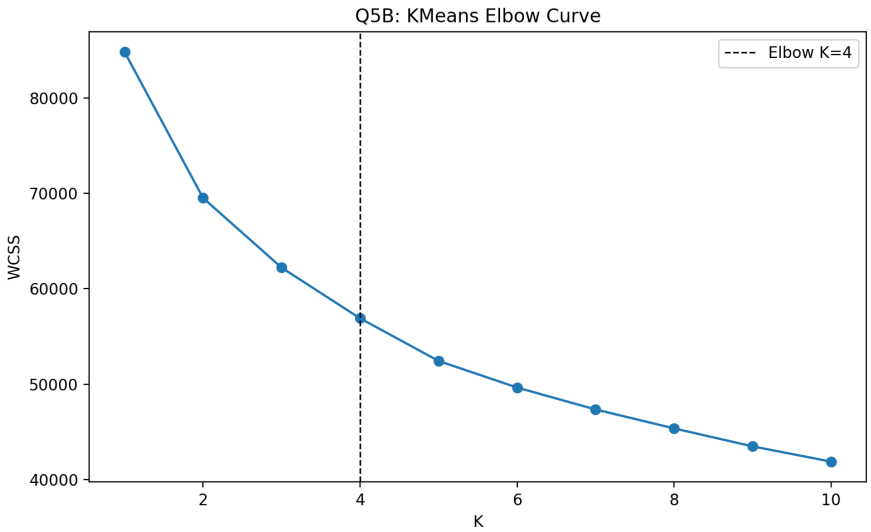
۶) یادگیری بدون نظارت

۱-۶) PCA

- EVR(PC1): ۰.۲۷۷۵
- EVR(PC2): ۰.۱۴۴۹
- EVR(PC1+PC2): ۰.۴۲۲۳

۲-۶) KMeans Elbow

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

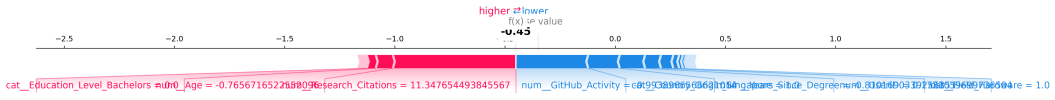


شکل ۸: تشخیص نقطه آرنج برای انتخاب تعداد خوشه‌ها.

۷) کپستون تبیین‌پذیری با SHAP

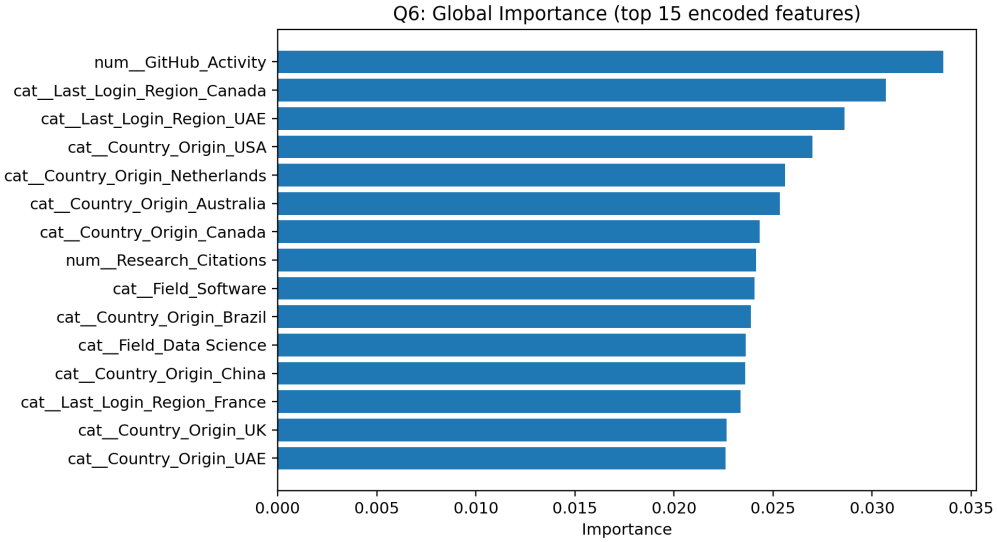
۷-۱) خروجی محلی نمونه

- 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.

۷-۲) نمای سراسری اهمیت ویژگی‌ها

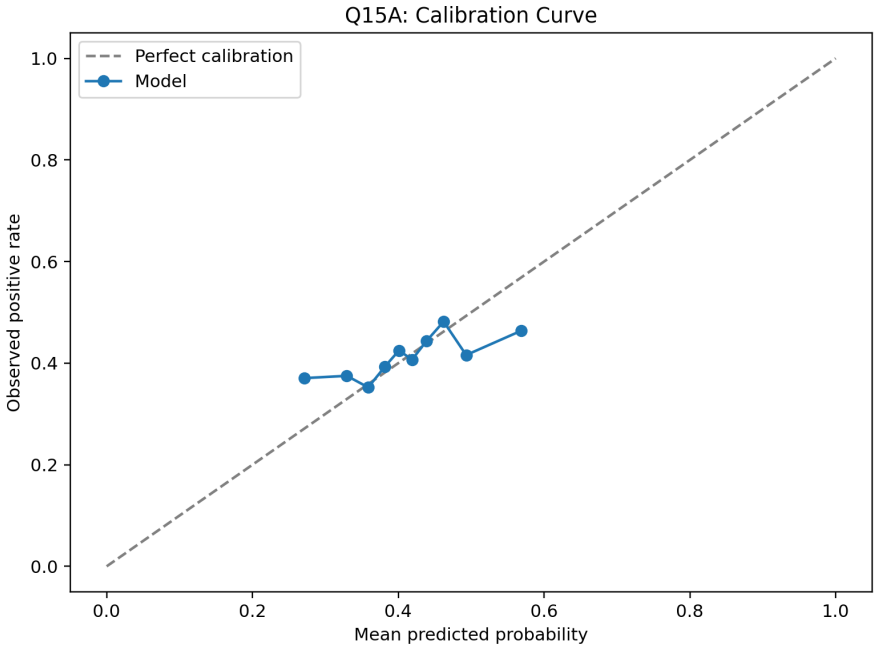


شکل ۱۰: اهمیت سراسری ویژگی‌ها بر اساس SHAP.

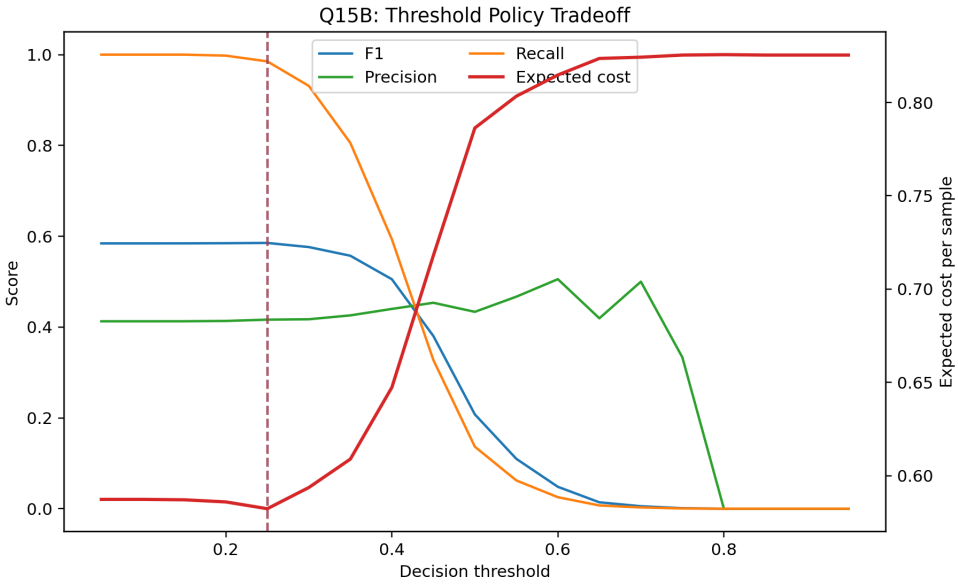
۸) کالیبراسیون و سیاست آستانه (Q۱۵)

کالیبراسیون احتمال و انتخاب آستانه برای مدل کپستون:

- Brier و ECE: مقادیر به‌روز در run_summary.json.
- دو آستانه گزارش می‌شود: یکی برای بیشینه‌سازی F1 و دیگری برای کمینه‌سازی هزینه نامتقارن خطا.



شکل ۱۱: منحنی کالیبراسیون مدل کپستون.

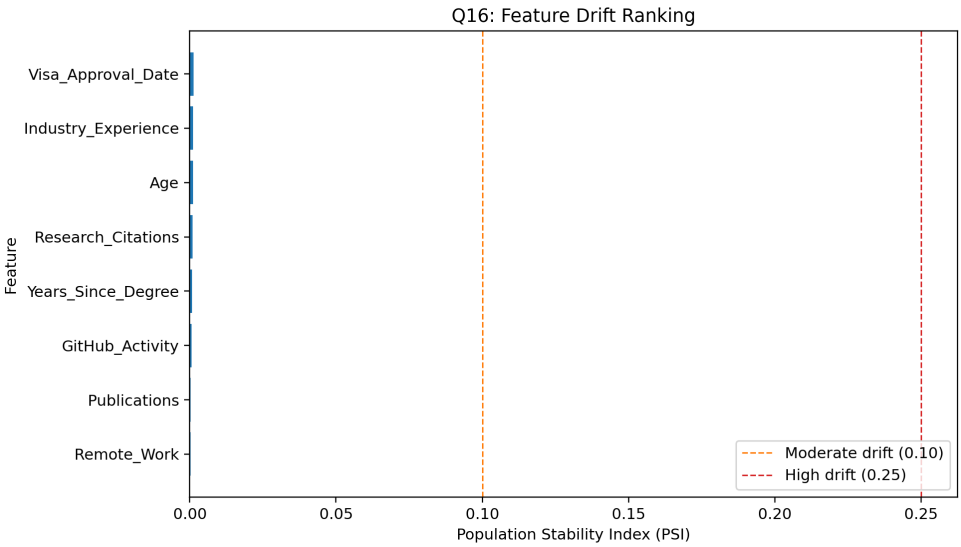


شکل ۱۲: تجارت آستانه: دقت/بازخوانی/اف۱ و هزینه مورد انتظار.

۹) درفت و پایداری داده (Q۱۶)

پایش درفت بین دو پنجره مرجع/جاری با PSI و یک شاخص دسته‌ای:

- قانون تقسیم (زمانی یا تصادفی) و اندازه هر پنجره در run_summary.json.
- رتبه‌بندی PSI برای ویژگی‌ها در code/solutions/q16_drift_psi.csv.
- شاخص JS divergence برای توزیع کشور (در صورت وجود).

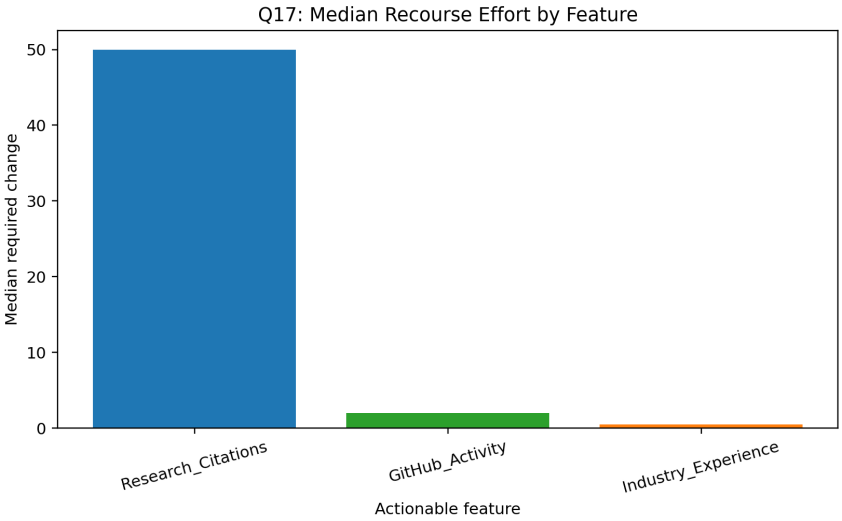


شکل ۱۳: ۱۲ ویژگی اول بر اساس PSI. خطوط عمودی: ۰.۱۰ (هشدار) و ۰.۲۵ (بحرانی).

۱۰) ریکورس مقابله‌ای (Q۱۷)

تحلیل کمینه تغییر لازم برای عبور از آستانه تصمیم برای نمونه‌های نزدیک مرز:

- نرخ موفقیت ریکورس و میانه تغییر برای ویژگی‌های عملیاتی در run_summary.json.
- جدول نمونه‌ها: code/solutions/q17_recourse_examples.csv



شکل ۱۴: میانه مداخله مورد نیاز برای هر ویژگی قابل اقدام.

۱۱) 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.

۱۲) 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

۱۳) محدودیت‌ها

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

۱۴) 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.