

گزارش نهایی کامل پروژه در قالب IEEE

تحلیل داده‌محور مهاجرت جهانی استعدادهای فنی

تیم دستیاران آموزشی درس علم داده
دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران
بسنۀ کپسون کارشناسی ارشد – بهار ۱۴۰۴

چکیده

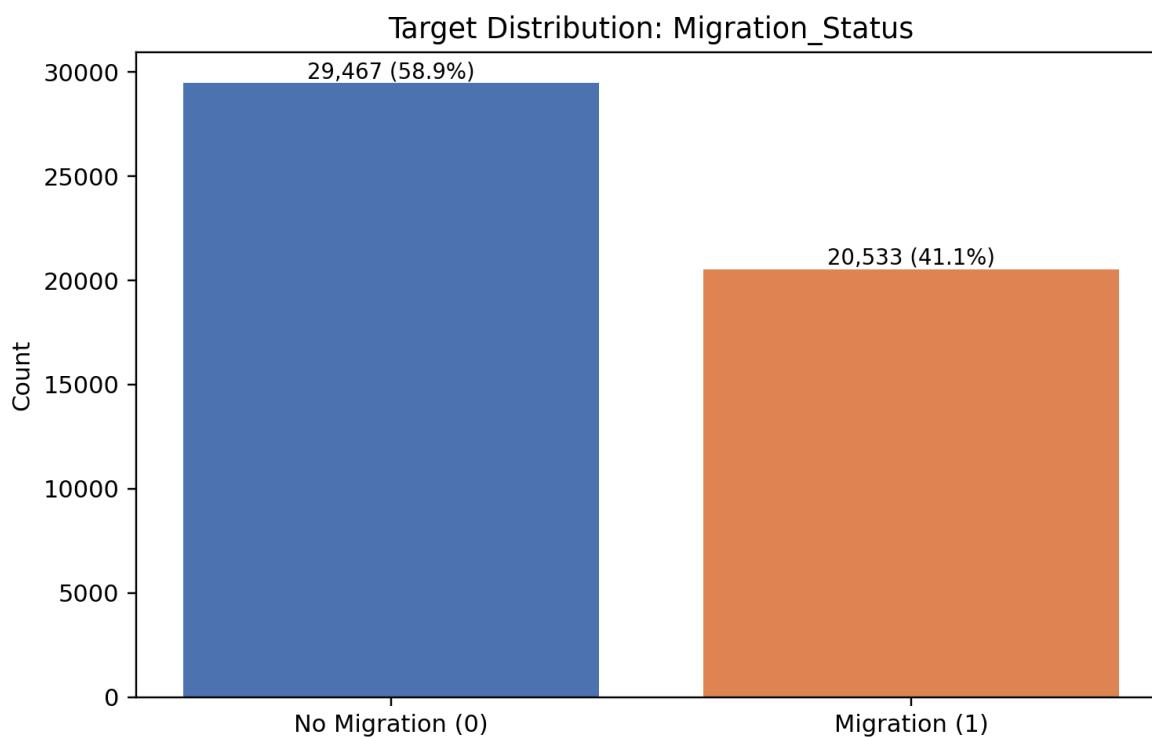
این گزارش نسخه کامل و قابل داوری پروژه نهایی درس علم داده است. پیاده‌سازی حاضر کل مسیر مهندسی داده تا تحلیل‌های تولیدی را پوشش می‌دهد: کنترل نشت، مدل‌سازی خطی و غیرخطی، بهینه‌سازی، یادگیری بدون ناظارت، تبیین‌پذیری با SHAP، و سه افزونه پیشرفته Q18-Q20 برای پایداری زمانی، عدم‌قطعیت و عدالت الگوریتمی.

تعريف مسئله و دامنه پروژه

هدف، پیش‌بینی Migration_Status برای ۵۰ هزار متخصص فناوری است. خروجی نهایی صرفاً یک مدل نیست؛ بلکه یک بسته کامل دانشگاهی-مهندسی است که شامل کد، آزمون، نوتبوک، گزارش دوزبانه و خروجی‌های استاندارد CSV/JSON/PNG است.

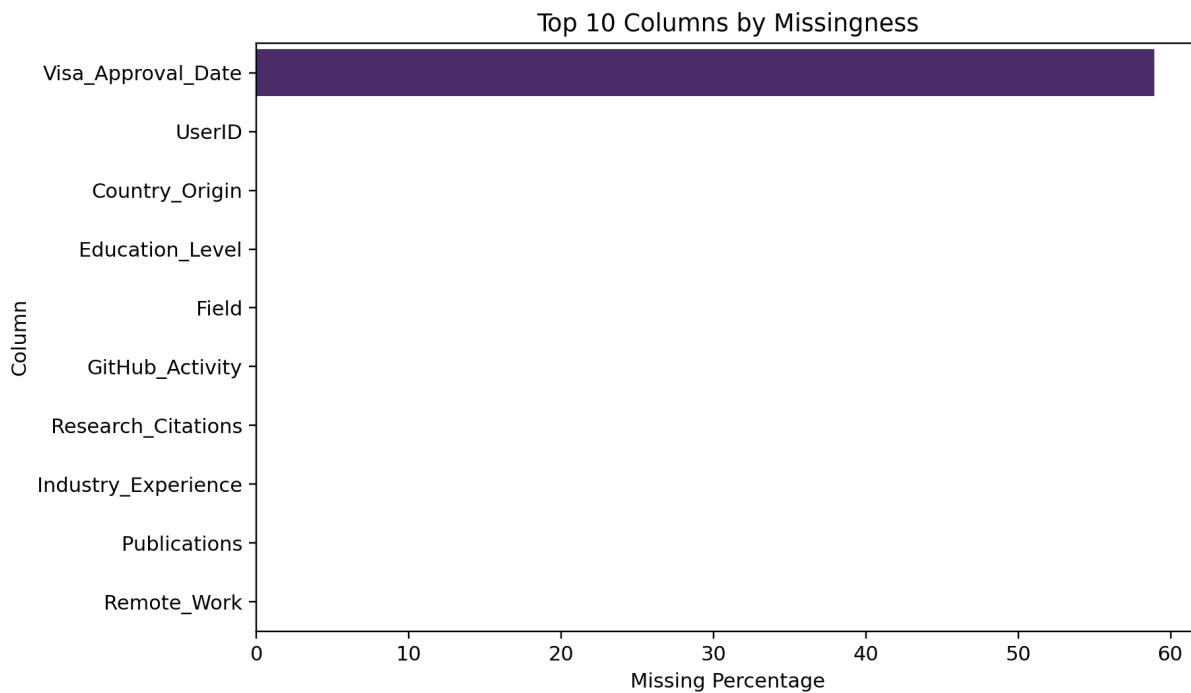
تشخیص اولیه داده

توازن کلاس هدف



شکل ۱: توزیع کلاس‌های متغیر هدف .Migration_Status

تفسیر: توزیع کلاس‌ها نامتوازن خفیف است و ارزیابی باید فراتر از دقت خام باشد.
اثر تصمیمی: معیارهای AUC/F1 و تحلیل آستانه، معیار اصلی انتخاب مدل قرار می‌گیرند.
حدودیت/تهدید: نسبت کلاس‌ها ممکن است در زمان استقرار تغییر کند و ثابت فرض کردن آن پربریسک است.
الگوی داده‌های گمشده



شکل ۲: ده ستون با بیشترین نرخ داده گمشده.

تفسیر: گمشده‌گی در ستون‌های فرایندی متمرکز است و بخشی از آن با نشت داده همپوشانی دارد.
اثر تصمیمی: ویژگی‌های پسربخداد، قبل از آموزش حذف می‌شوند یا بهصورت کنترل شده وارد تحلیل می‌شوند.
حدودیت/تهدید: اگر گمشده‌گی تصادفی نباشد، خود الگوی گمشده‌گی می‌تواند منشا بایاس شود.

ساختار همبستگی

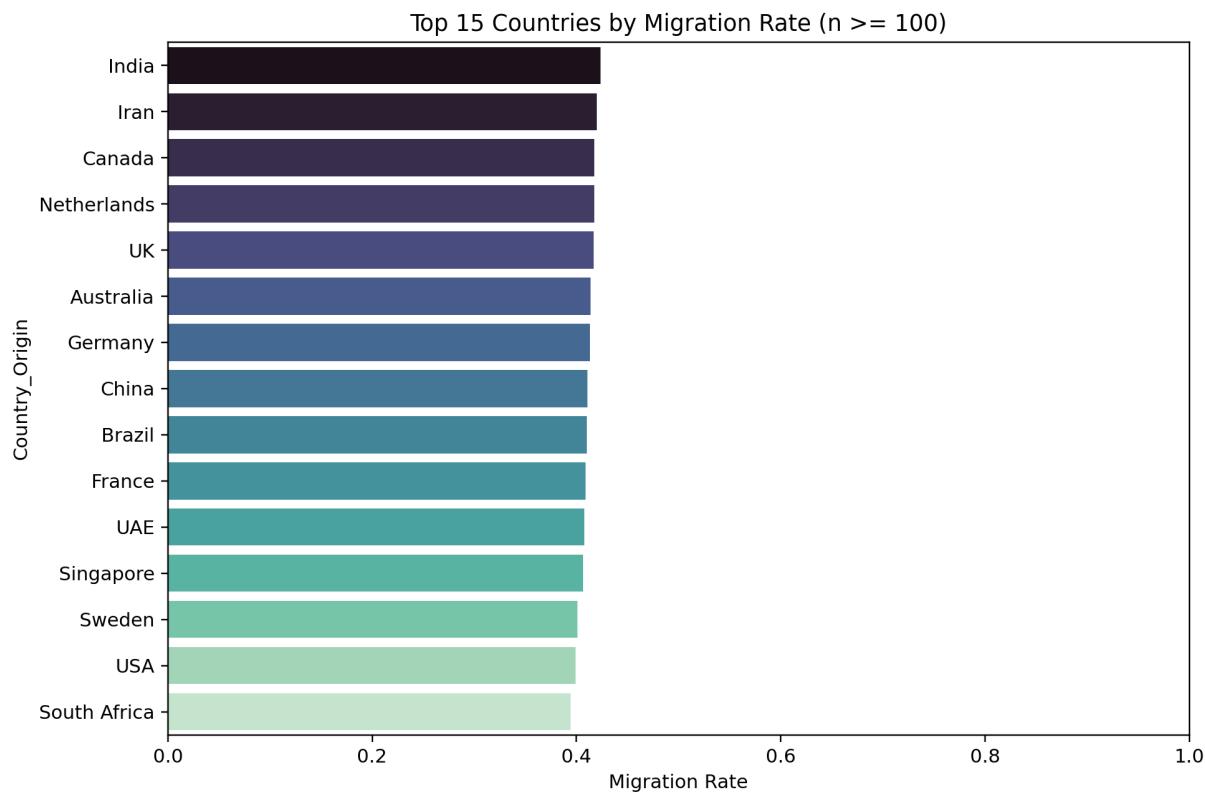


Correlation Heatmap (Target + Top Numeric Features)								
Migration_Status	1.00	0.12	0.05	0.01	0.01	0.00	-0.00	-0.00
GitHub_Activity	0.12	1.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00
Research_Citations	0.05	-0.00	1.00	-0.00	-0.01	-0.00	-0.01	-0.00
Remote_Work	0.01	-0.00	-0.00	1.00	0.00	0.00	0.00	0.00
Publications	0.01	-0.00	-0.01	0.00	1.00	0.01	0.00	0.01
Years_Since_Degree	0.00	-0.00	-0.00	0.00	0.01	1.00	-0.00	0.94
Industry_Experience	-0.00	0.00	-0.01	0.00	0.00	-0.00	1.00	-0.00
Age	-0.00	-0.00	-0.00	0.00	0.01	0.94	-0.00	1.00

شکل ۳: نقشه همبستگی ویژگی‌های عددی کلیدی با هدف.

تفسیر: چند ویژگی مهم سیگنال دارند اما هیچ متغیری به تنهایی پاسخ کامل نمی‌دهد.
اثر تصمیمی: استفاده از مدل‌های چندمتغیره و غیرخطی به صورت همزمان توجیه می‌شود.
حدودیت/تهذیب: همبستگی دلالت علی ندارد و می‌تواند بازتاب عوامل پنهان سیاستی باشد.

نرخ مهاجرت در سطح کشور



شکل ۴: مقایسه نرخ مهاجرت بین کشورها پس از اعمال حداقل حجم نمونه.

تفسیر: تفاوت گروهی بین کشورها معنادار است و بیتوجهی به آن ارزیابی را ناقص می‌کند.

اثر تصمیمی: تحلیل عدالت قبل و بعد از مداخله الزامی است.

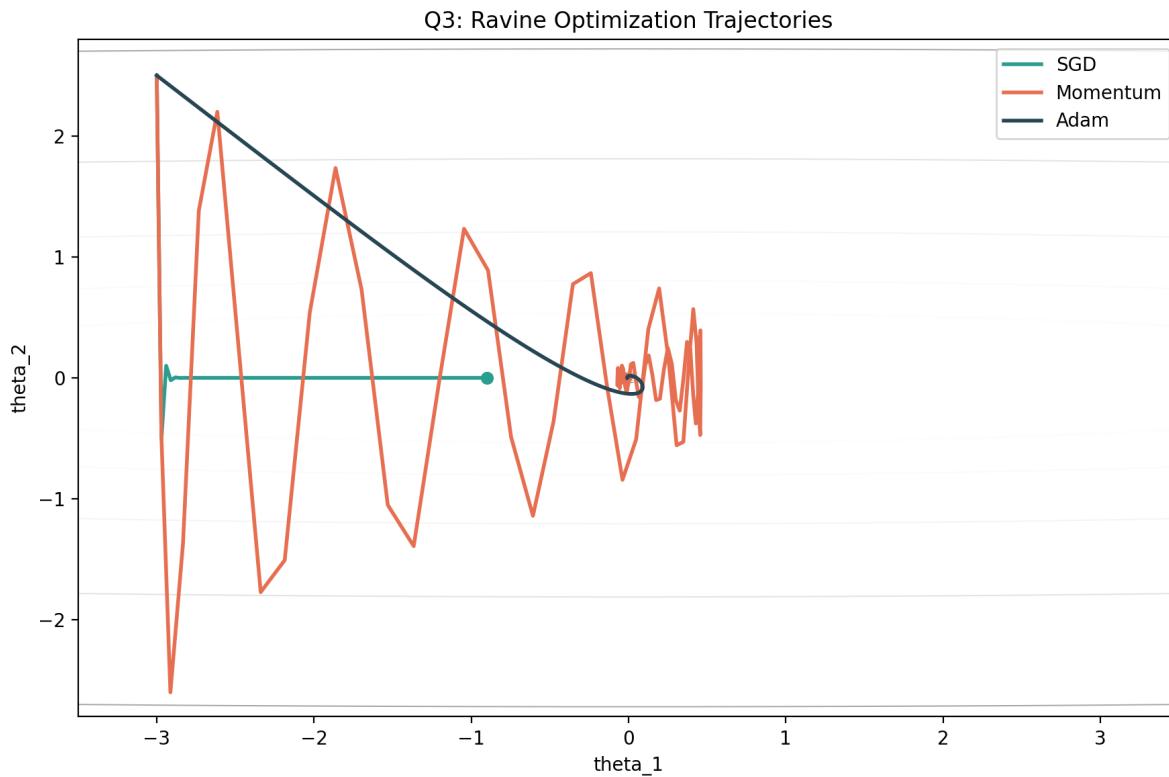
حدودیت/تهدید: تفاوت‌ها می‌توانند ناشی از محدودیت‌های سیاستی باشند نه توان واقعی افراد.

نتایج هسته اصلی (Q1) تا (Q6)

Q1: مهندسی داده و نشت

خروجی SQL در code/solutions/q1_moving_average.sql ذخیره شده است. ویژگی Visa_Approval_Date به عنوان نشت مستقیم حذف می‌شود.

Q3: تحلیل بهینه‌سازها



شکل ۵: مسیر همگرایی Momentum، SGD و Adam روی سطح دره‌ای.

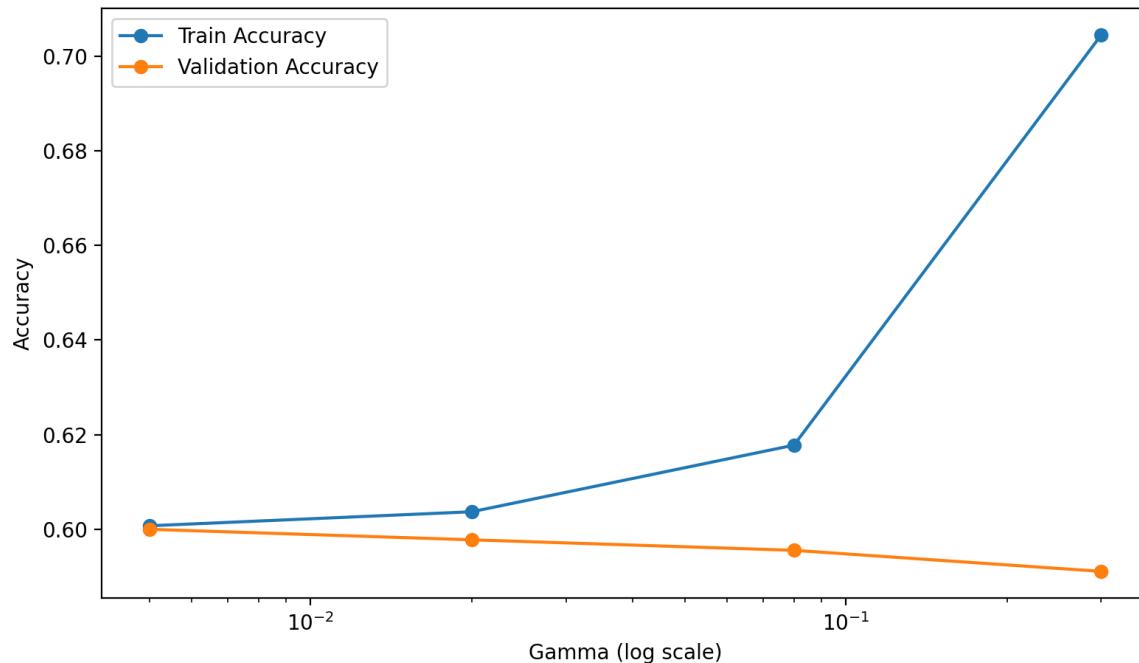
تفسیر: Momentum/Adam نوسان عرضی را کاهش می‌دهند و همگرایی سریع‌تری دارند.

اثر نصیبی: برای توابع بذرگانه، بهینه‌ساز شتابدار انتخاب پیش‌فرض مناسب‌تری است.

حدودیت/تهذید: رفتار روی تابع آزمایشی، کل پیچیدگی اهداف غیرمحدب را پوشش نمی‌دهد.

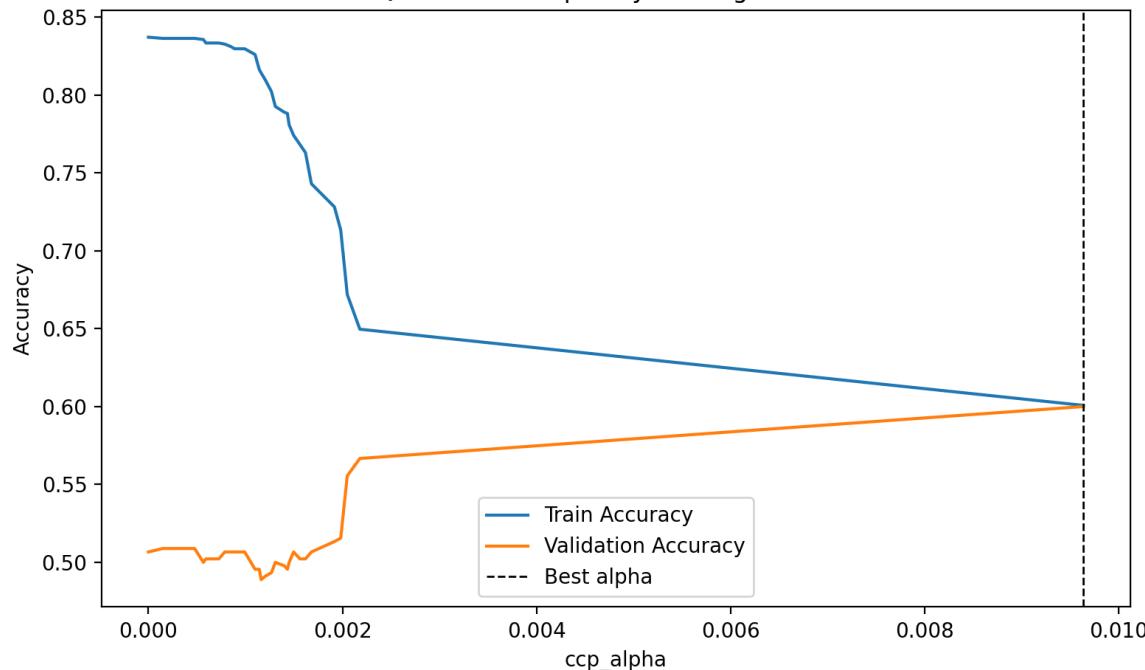
Q4: مدل غیرخطی و کنترل پیچیدگی

Q4A: SVM RBF Gamma vs Performance

شکل ۶: حساسیت دقت اعتبارسنجی نسبت به γ در SVM-RBF.

تفسیر: افزایش γ می‌تواند مرز تصمیم را بیش از حد محلی و پرنوسان کند.
اثر تصمیمی: در حالت بیش‌برازش، γ کاهش می‌یابد و روی اعتبارسنجی تنظیم می‌شود.
حدودیت/تهذیب: نتیجه نهایی به مقیاس‌بندی و همپوشانی کلاس‌ها وابسته است.

Q4B: Cost-Complexity Pruning Tradeoff

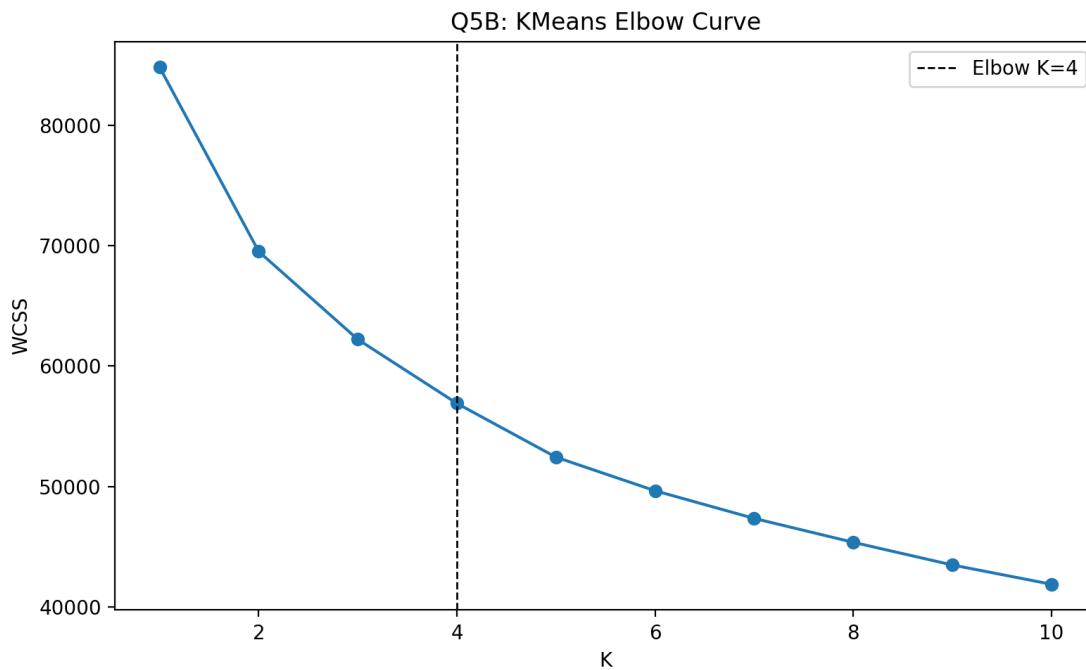


شکل ۷: مصالحه خط‌پیچیدگی در هرس Decision Tree

تفسیر: α بزرگتر اندازه درخت را کم می‌کند و واریانس را پایین می‌آورد.
اثر تصمیمی: انتخاب α بر اساس عملکرد اعتبارسنجی انجام می‌شود، نه خطای آموختش.

محدودیت/تهدید: منحنی هرس نسبت به نحوه تقسیم داده حساس است.

Q5: ساختار بدون نظارت



شکل ۸: منحنی آرنج WCSS برای انتخاب تعداد خوشه‌ها.

تفسیر: پس از یک K میانی، کاهش WCSS بازده نزولی پیدا می‌کند.

اثر تصمیمی: K به صورت heuristic انتخاب و سپس با تفسیرپذیری خوشه‌ها اعتبارسنجی می‌شود.

محدودیت/تهدید: در منحنی‌های کم‌اچنا، محل آرنج می‌تواند مبهم باشد.

Q6: مدل کپستون و SHAP

پروفایل اجرا: balanced

مدل کپستون: XGBoost

AUC: ۰.۵۴۹۵، F1: ۰.۵۸۳۵، Accuracy: ۰.۲۴۷۵

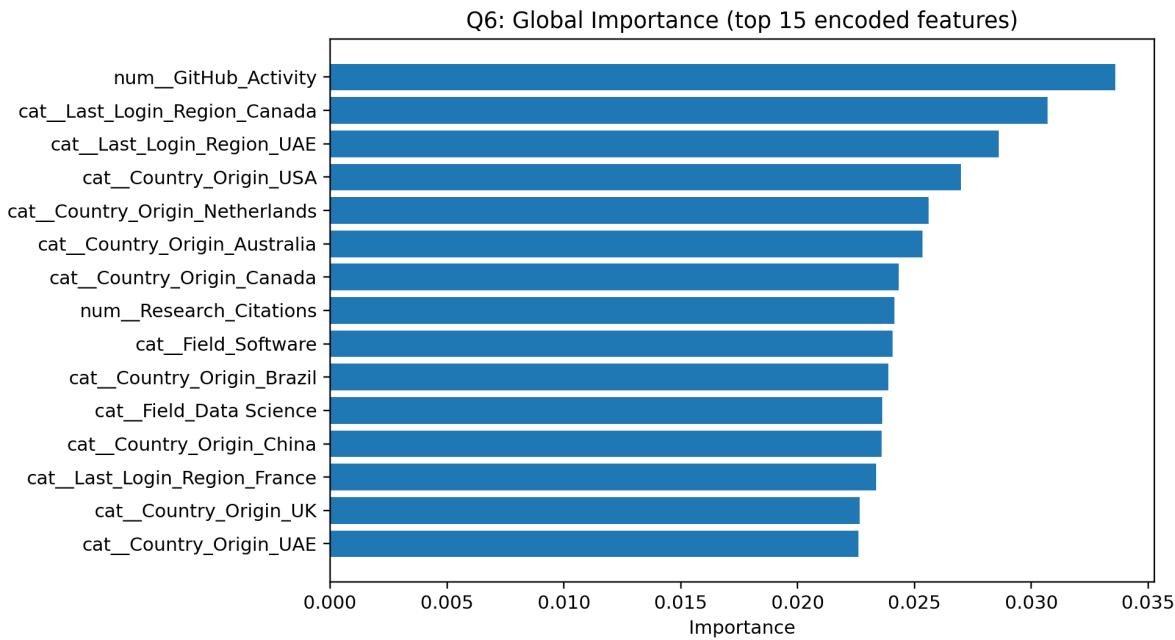


شکل ۹: توضیح محلی SHAP برای نمونه منتخب.

تفسیر: اختلاف base value و خروجی نهایی از جمع سهم ویژگی‌ها ساخته می‌شود.

اثر تصمیمی: بازبینی پرونده روی عوامل غالب منفی/مثبت مرکز می‌شود.

محدودیت/تهدید: SHAP رفتار مدل را توضیح می‌دهد، نه رابطه علی‌دنیای واقعی را.



شکل ۱۰: نمای سراسری اهمیت ویژگی‌ها در مدل کپستون.

تفسیر: ویژگی‌های مرتبط با فعالیت فنی و استنداد، وزن بیشتری در پیش‌بینی دارند.

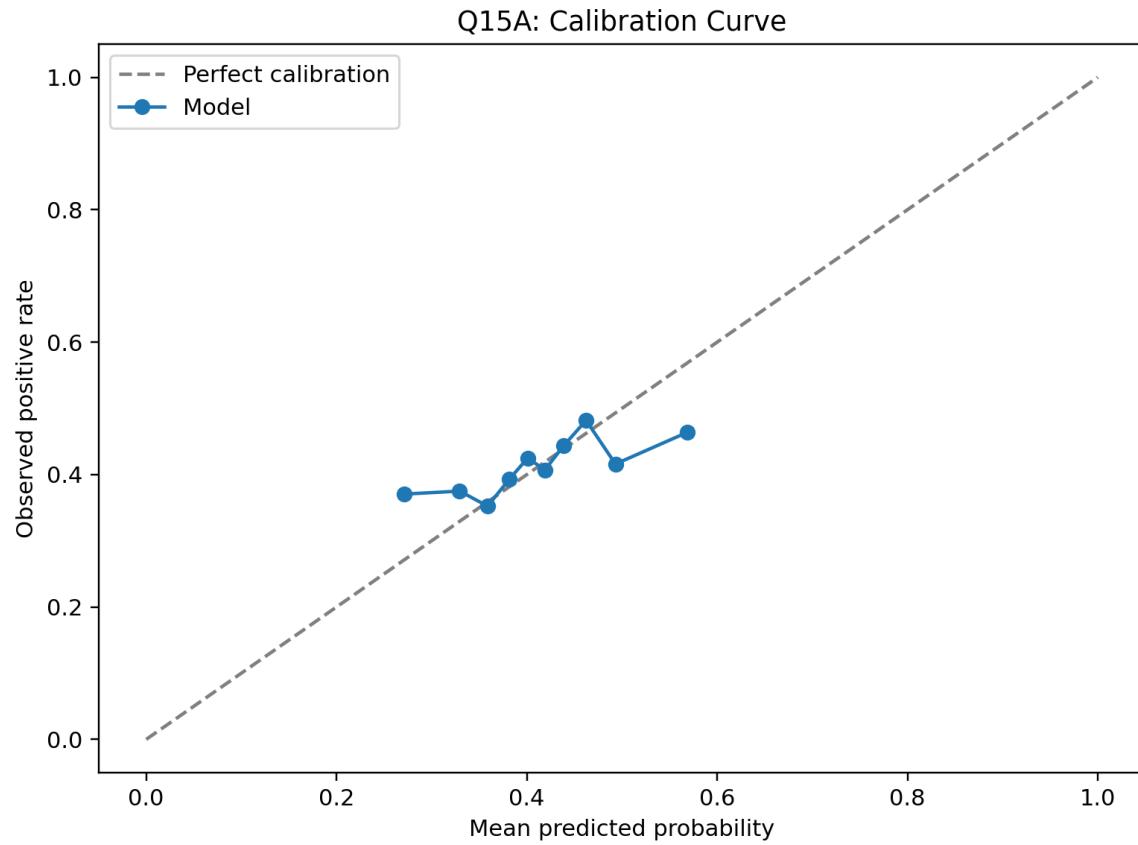
اثر تصمیمی: کنترل کیفیت داده برای ویژگی‌های پراثر، اولویت حاکمیتی می‌گیرد.

حدودیت/تهذید: اهمیت سراسری، ناهمگنی اثر در همه زیرگروه‌ها را نشان نمی‌دهد.

بلوک پیشرفت‌هه تولیدی (Q15) تا (Q20)

Q15: کالیبراسیون و سیاست آستانه

ECE: ۲۴۳۶.۰، F1: ۲۵۰۰.۰، آستانه بینه: ۰۳۲۷.۰

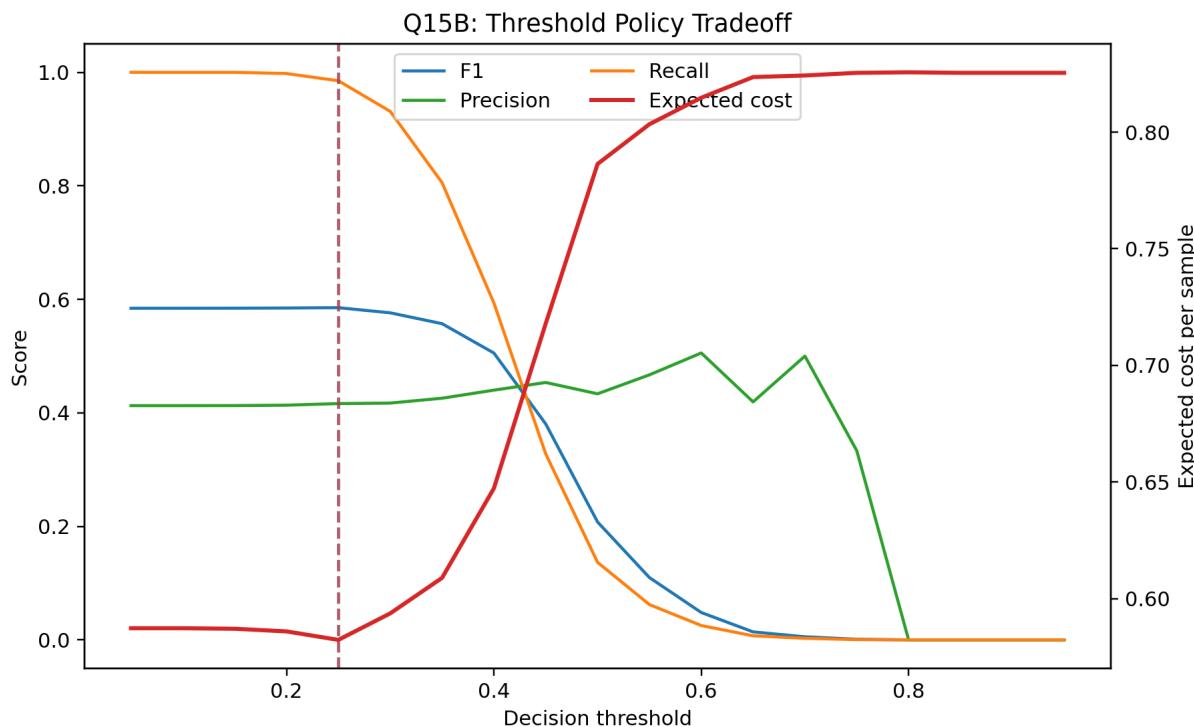


شکل ۱۱: منحنی قابلیت اطمینان کالیبراسیون احتمال.

تفسیر: کالیبراسیون نشان می‌دهد احتمال پیش‌بینی شده تا چه حد با فراوانی واقعی هم‌است است.

اثر تصمیمی: تصمیم‌های حساس با احتمال کالیبره شده گرفته می‌شوند، نه امتیاز خام.

حدودیت/تهدید: کالیبراسیون در گذر زمان ممکن است فرسوده شود.

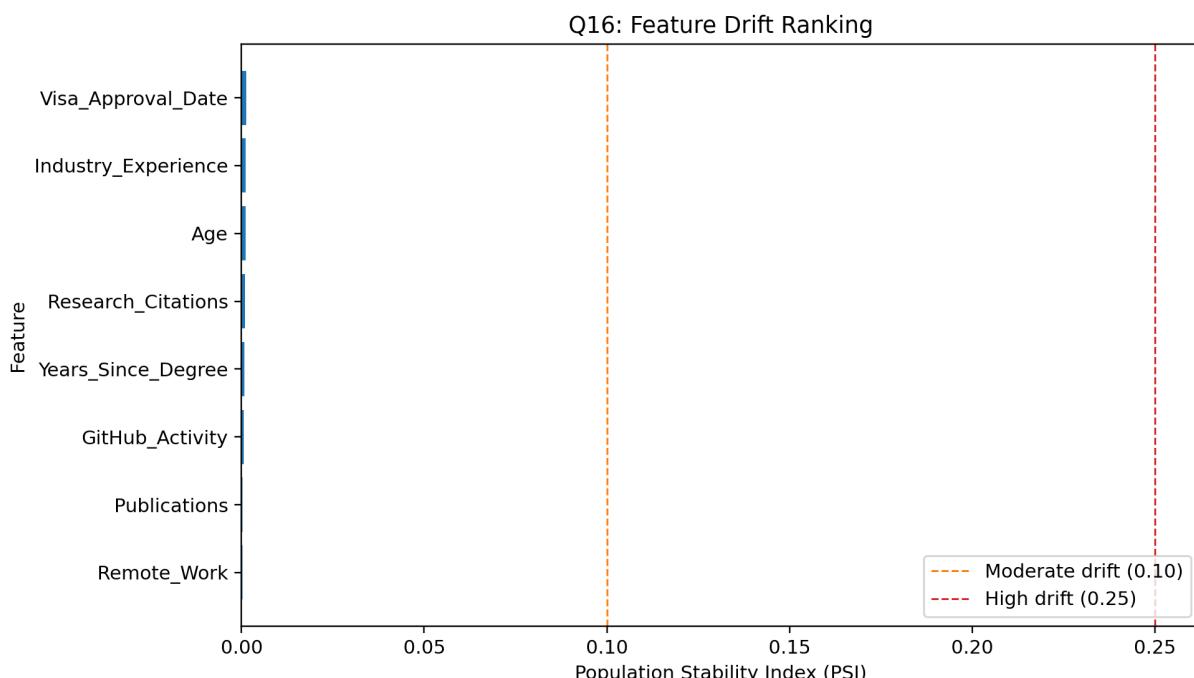


شکل ۱۲: مصالحه آستانه بین دقت، بازخوانی، F1 و هزینه مورد انتظار.

تفسیر: هر آستانه، توازن متفاوتی بین خطاها نوع اول/دوم ایجاد می‌کند.
اثر تصمیمی: آستانه نهایی با ماتریس هزینه سیاستی انتخاب می‌شود.
حدودیت/تهذید: مفروضات هزینه ممکن است بین کشور/سازمان متفاوت باشد.

Q16: پایش درفت

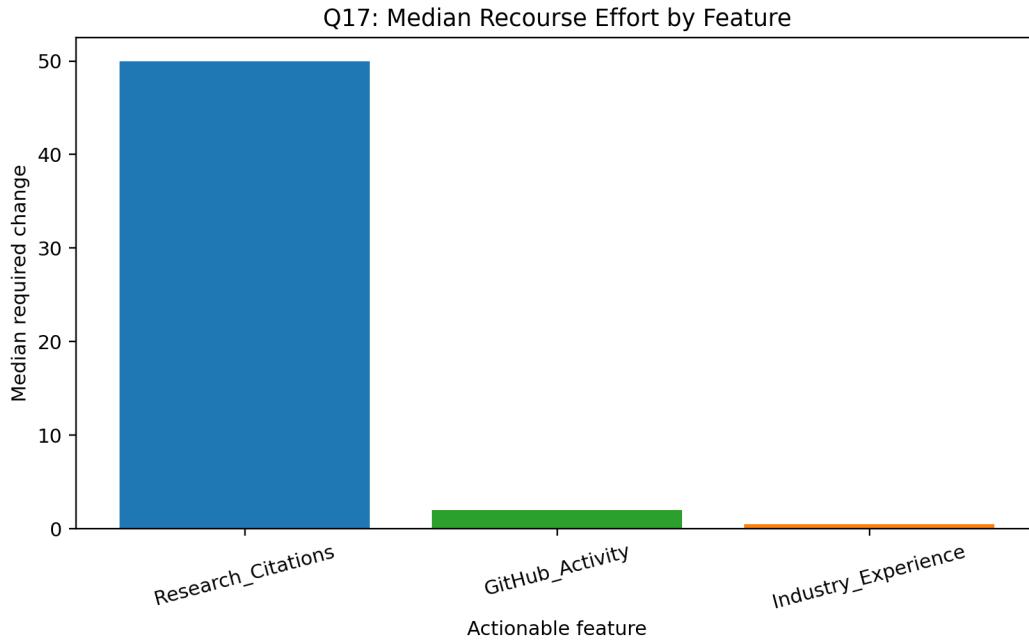
ویژگی با بیشترین درفت: $\text{Visa_Approval_Date} = \text{PSI} + 0.13$



شکل ۱۳: رتبه‌بندی درفت ویژگی‌ها بر اساس PSI.

تفسیر: بخشی از ویژگی‌ها ناپایداری توزیعی قابل‌توجه دارند.
اثر تصمیمی: استانه‌های هشدار PSI به پایش دوره‌ای و بازآموزی متصل می‌شود.
حدودیت/تهدید: PSI تغییر رابطه ویژگی-هدف را مستقیماً اندازه نمی‌گیرد.

Q17: ریکورس مقابله‌ای
نرخ موفقیت ریکورس: .۰۰۰۰۱

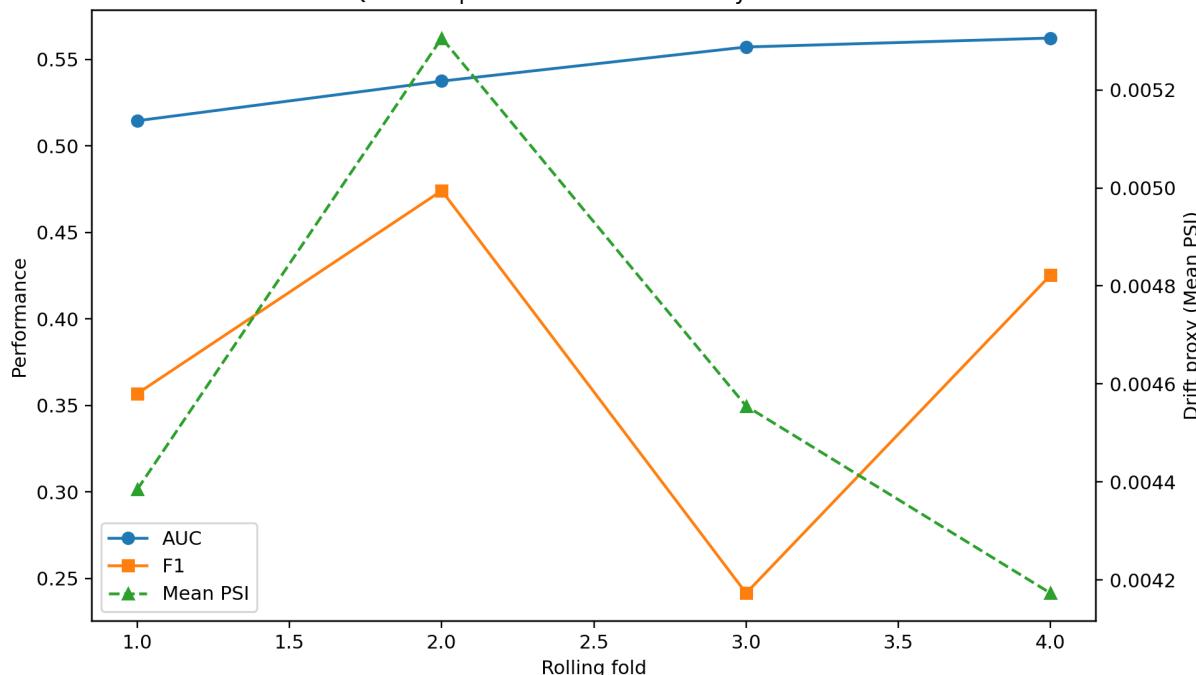


شکل ۱۴: میانه تغییر لازم برای عبور از مرز تصمیم در هر ویژگی قابل اقدام.

تفسیر: هزینه تغییر در ویژگی‌های عملیاتی پکسان نیست و قابل کمی‌سازی است.
اثر تصمیمی: پیشنهاد اقدام به متقاضی بر اساس کم‌هزینه‌ترین مسیر ممکن داده می‌شود.
حدودیت/تهدید: عملی‌بودن ریکورس به حدودیت‌های واقعی خارج از داده وابسته است.

Q18: اعتبارسنجی زمانی و افت عملکرد
میانگین AUC زمانی: .۵۴۲۸۰۰، افت AUC: .۰۴۷۸۰۰

Q18: Temporal Performance Decay vs Drift



شکل ۱۵: افت عملکرد در پنجره‌های زمانی غلطان در کنار شاخص درفت.

تفسیر: عملکرد بین های fold متوالی تغییر می‌کند و با شدت درفت قابل مقایسه است.

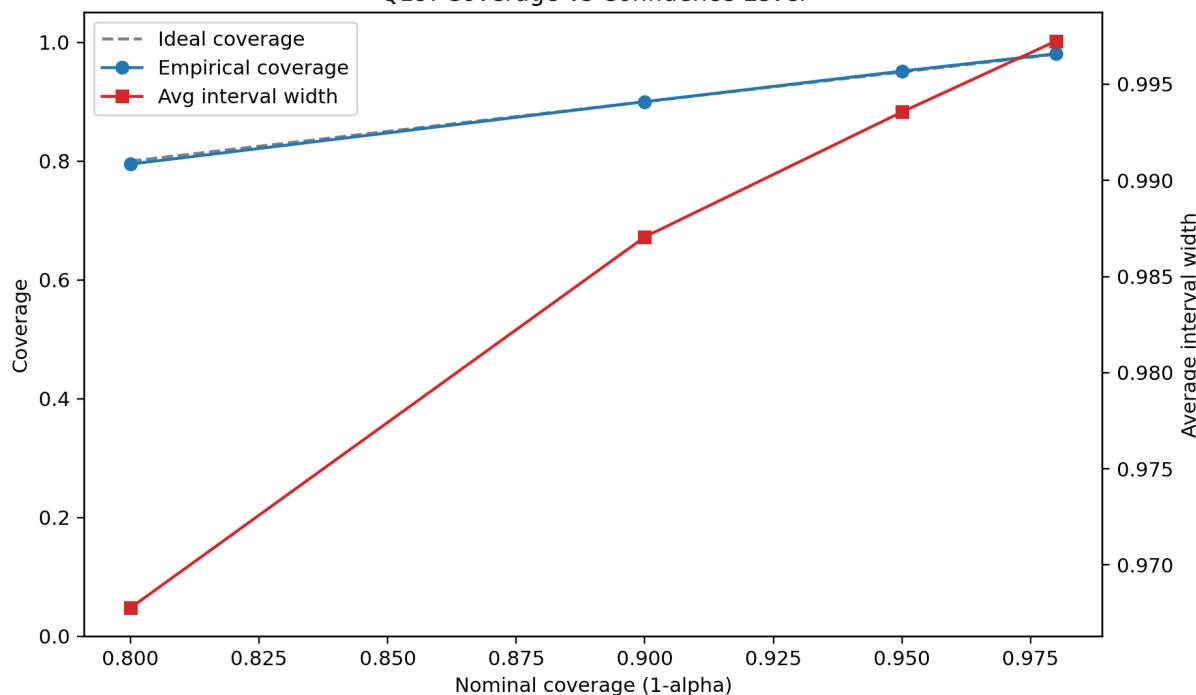
اثر نصیبی: قبل از استقرار، سنجش زمانی اجباری است و صرف random split کافی نیست.

حدودیت/تهدید: در نبود ستون زمانی معتبر، fallback باید صریح گزارش شود.

Q19: کمی‌سازی عدمقطعیت

Coverage@90 = ۹۰۰۰۰: بیشینه کمپوشنی

Q19: Coverage vs Confidence Level



شکل ۱۶: پوشش اسمی در برابر پوشش تجربی و پهنای بازه عدمقطعیت.

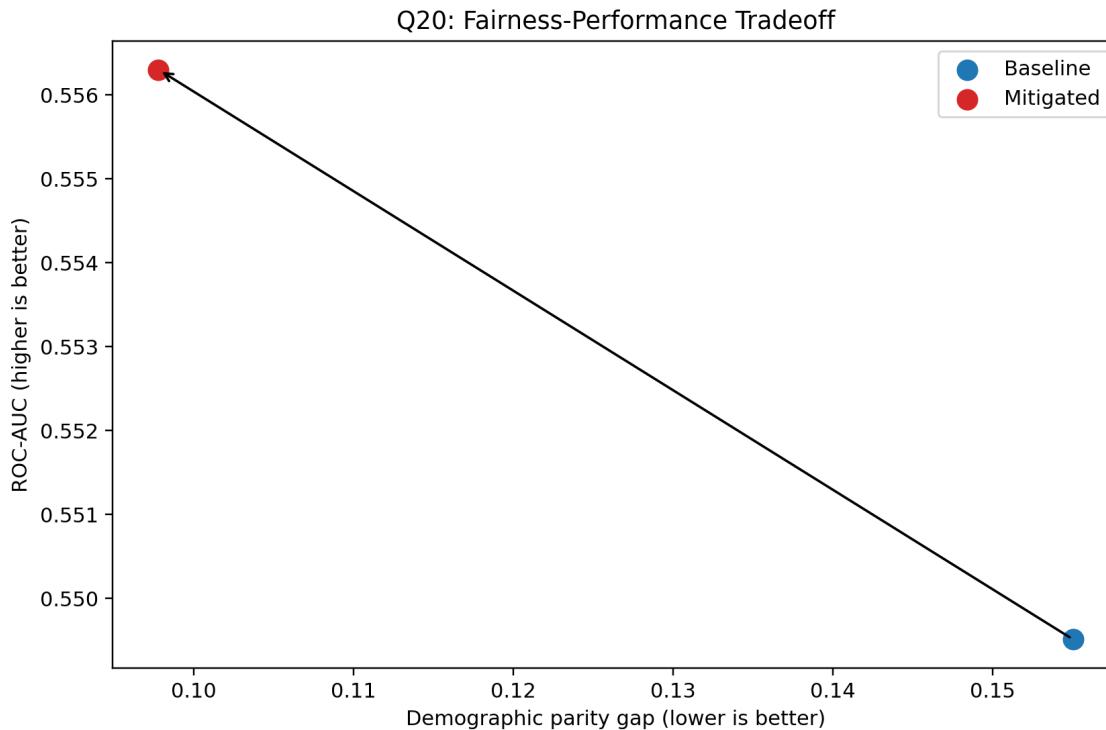
تفسیر: بازده‌های conformal توازن بین اطمینان پیش‌بینی و عرض بازه را روشن می‌کنند.

اثر تصمیمی: موارد کم‌اطمینان به مسیر بررسی انسانی ارجاع می‌شوند.

حدودیت/تهدید: تصمین‌های پوشش در صورت جایه‌جایی توزیع تصنیف می‌شوند.

Q20: مداخله عدالت‌الگویی

شکاف DP پایه: ۱۵۵۰.۰، شکاف DP پس از مداخله: ۰۹۷۸.۰، نتیجه قید سیاستی: **true**.



شکل ۱۷: مصالحه عدالت-عملکرد از مدل پایه تا مدل مداخله شده.

تفسیر: بازوزن‌دهی، نقطه عملکردی مدل را روی صفحه عدالت-کارایی جایه‌جا می‌کند.

اثر تصمیمی: استقرار منوط به عبور همزمان از قید افت عملکرد و بهبود عدالت است.

حدودیت/تهدید: بهبود یک معیار عدالت ممکن است برای زیرگروه‌های ریزتر کافی نباشد.

بازتولیدنی و خروجی‌ها

اجرای کامل با دستور زیر انجام می‌شود:

```
python code/scripts/full_solution_pipeline.py --profile {fast,balanced,heavy}
latex_metrics
code/solutions نخیره می‌شوند، شامل run_summary.json نسخه ۲، خروجی‌های Q18-Q20 و فایل‌های
code/figures خروجی‌ها در برای گزارش خودکار.
```

ضمیمه A: تشریح کدهای پژوهه

نقشه مأذول‌ها

full_solution_pipeline.py ارکستریتور مرکزی Q1-Q20 است و جریان داده، مدل، خروجی و گزارش را یکپارچه می‌کند. q18_temporal.py اعتبارسنجی زمانی غلطان، افت AUC/F1 و شاخص درفت را تولید می‌کند. q19_split-conformal.py بازه‌های split-conformal و تحلیل پوشش تجربی را می‌سازد. q20_fairness_mitigation.py مقایسه قبل/بعد مداخله عدالت و قید سیاستی استقرار را اجرا می‌کند. report_metrics_export.py متريک‌ها را به JSON/TEX تبدیل می‌کند تا گزارش‌ها hard-code نباشند. report_stats.py شکل‌های گزارش مدیریتی و report_assets.py را بازتولید می‌کند.

منطقه مهندسی

پایپ لاین با پروفایل‌های fast/balanced/heavy کنترل بودجه اجرا را انجام می‌دهد و در هر اجرا: ابتدا ویژگی‌های نشتزا حذف می‌شوند، سپس خروجی‌های تحلیلی CSV/PNG و در انتهای run_summary.json نسخه ۲ ساخته می‌شود. این طراحی باعث می‌شود ارزیابی علمی، گزارش LaTeX و تست خودکار همگی روی یک منبع حقیقت یکسان همگام باشند.

ضمیمه B: شناسنامه کامل شکل‌ها

شکل‌های گزارش پایه داده

report_target_balance.png: هدف، بررسی توازن کلاس است. اگر عدم توازن زیاد باشد، دقت خام گمراهنده می‌شود و باید از AUC/F1 استفاده شود.

report_missingness_top10.png: ستون‌های دارای گمشدگی بالا را نشان می‌دهد و ورودی تصمیم‌های حذف/جاوگزینی است. همبستگی اولیه ویژگی‌ها با هدف را نمایش می‌دهد و برای اولویت‌بندی تحلیل چندمتغیره استفاده می‌شود.

report_numeric_correlation.png: تفاوت نرخ مهاجرت بین کشورها را نشان می‌دهد و ضرورت تحلیل عدالت را تقویت می‌کند.

شکل‌های مدل‌سازی و یادگیری

q3_ravine_optimizers.png: مقایسه مسیر همگرایی SGD/Momentum/Adam روى تابع دره‌ای؛ مبنای انتخاب بهینه‌ساز پایدارتر است.

q4_svm_gamma_sweep.png: حساسیت SVM-RBF نسبت به γ و کنترل overfit را نشان می‌دهد.

q4_tree_pruning_curve.png: اثر tradeoff ccp_alpha بر ccp_alpha در درخت تصمیم.

q5_kmeans_elbow.png: منطق انتخاب K بر پایه بازده نزولی کاهش WCSS.

q6_shap_force_plot.png: تبیین محلی علت پیش‌بینی یک نمونه خاص.

q6_shap_summary.png: اهمیت سراسری ویژگی‌ها برای حاکمیت داده و کنترل کیفیت.

شکل‌های بلوک پیشرفته تولیدی

q15_calibration_curve.png: هماراستایی احتمال پیش‌بینی با نرخ واقعی رخداد؛ مبنای تصمیم‌های ریسک‌محور.

q15_threshold_tradeoff.png: مصالحه آستانه بین Precision/Recall/F1 و هزینه مورد انتظار.

q16_drift_psi_top12.png: رتبه‌بندی درفت ویژگی‌ها با PSI و تعریف تریگر بازارآمدی.

q17_recourse_median_deltas.png: میانه تغییر لازم برای recourse در ویژگی‌های قابل اقدام.

q18_temporal_degradation.png: افت زمانی عملکرد در کنار proxy درفت؛ معیار آمادگی استقرار زمانی.

q19_coverage_vs_alpha.png: پوشش اسمی-تجربی و پنهانی بازه عدم قطعیت؛ مبنای review human برای موارد کم‌اطمینان.

q20_fairness_tradeoff.png: جابه‌جایی نقطه مدل روی صفحه عدالت-کارایی پس از mitigation و ارزیابی قید سیاستی.

شکل تكمیلی آموزشی

خروجی نسخه baseline (legacy) آموزشی در train_and_explain.py است و جایگزین تحلیل کپی‌تون Q6 نیست.

جمع‌بندی

پروژه در این نسخه به یک بسته حرفه‌ای کپی‌تون تبدیل شده است: پیچیدگی علمی، استاندارد مهندسی، و قابلیت داوری آموزشی را بهصورت یکپارچه و قابل بازتولید پوشش می‌دهد.