

# 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

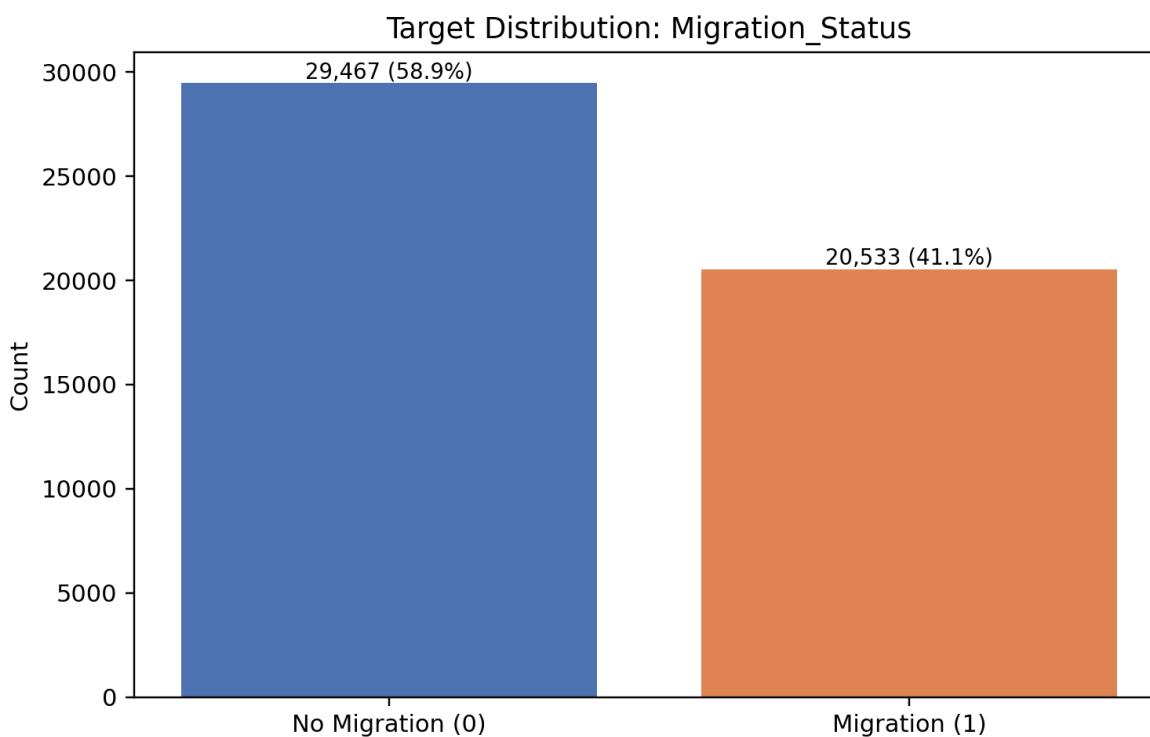
این گزارش نسخه کامل و قابل داوری پروژه نهایی درس علم داده است. پیاده‌سازی حاضر کل مسیر مهندسی داده تا تحلیل‌های تولیدی را پوشش می‌دهد: کنترل نشت، مدل‌سازی خطی و غیرخطی، بهینه‌سازی، یادگیری بدون ناظارت، تبیین‌پذیری با SHAP، و سه افزونه پیشرفته Q18-Q20 برای پایداری زمانی، عدم‌قطعیت و عدالت الگوریتمی.

### 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.

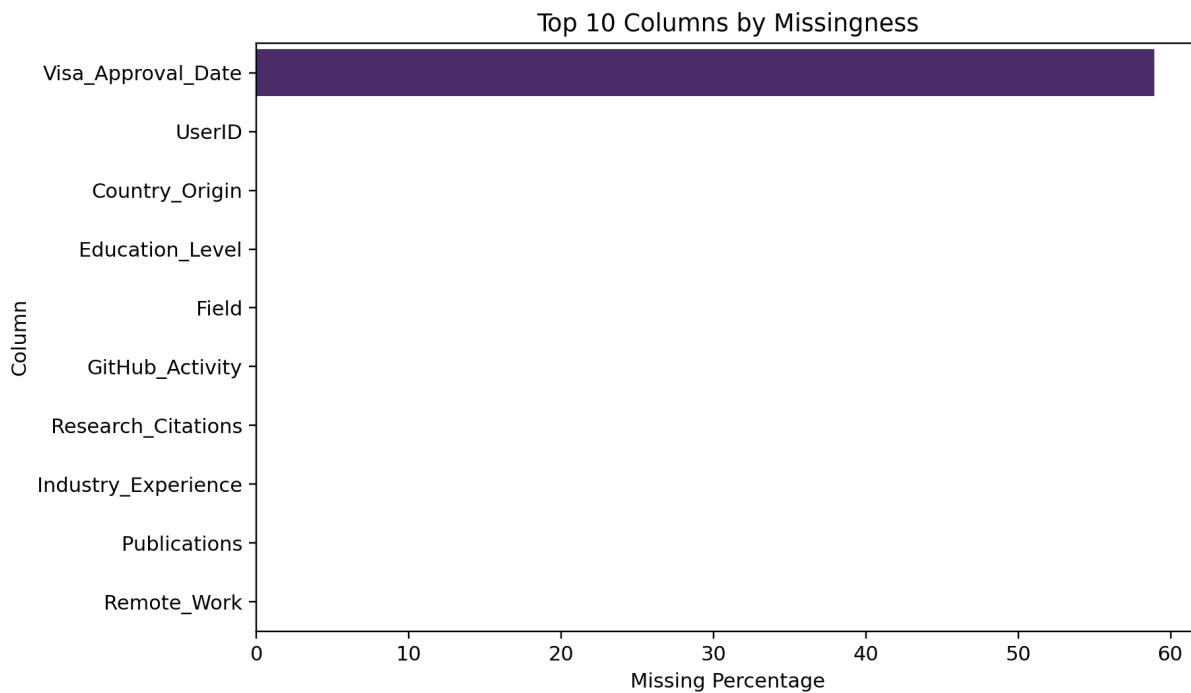
### تشخیص اولیه داده

### 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

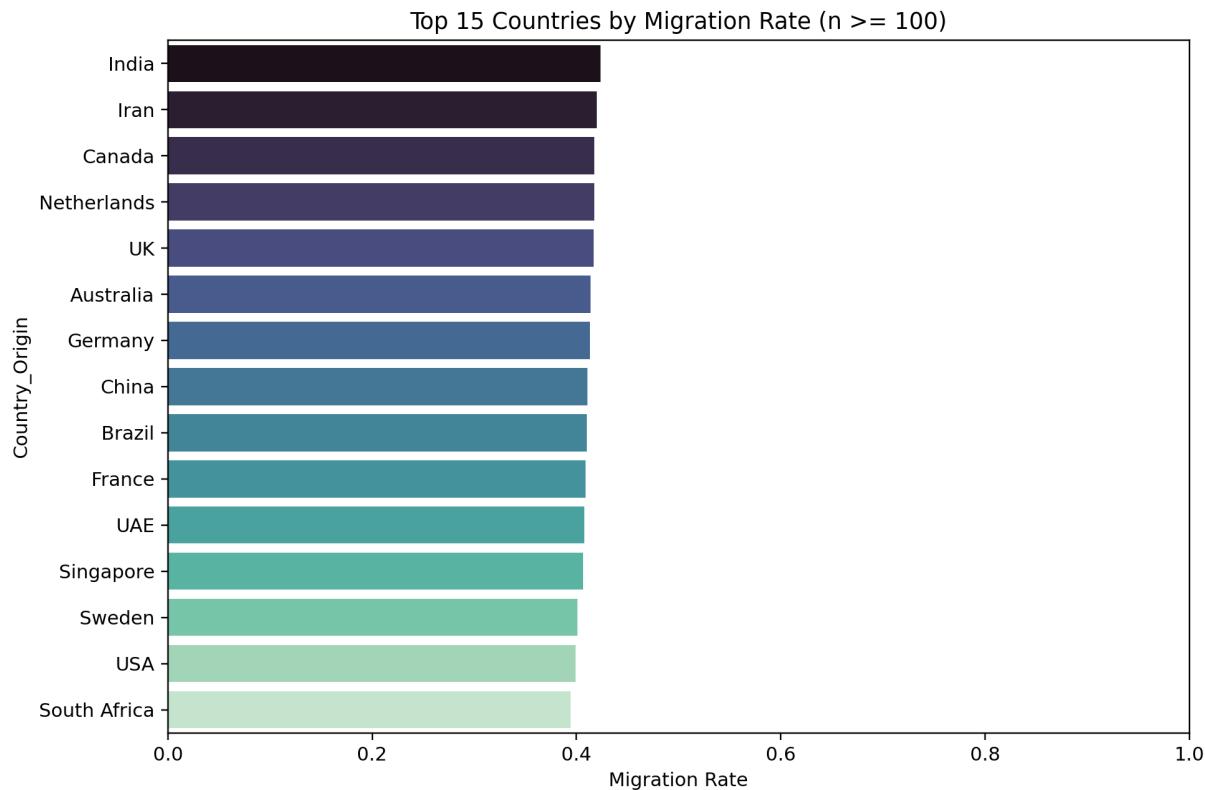


Correlation Heatmap (Target + Top Numeric Features)								
Migration_Status	1.00	0.12	0.05	0.01	0.01	0.00	-0.00	-0.00
GitHub_Activity	0.12	1.00	-0.00	-0.00	-0.00	-0.00	0.00	-0.00
Research_Citations	0.05	-0.00	1.00	-0.00	-0.01	-0.00	-0.01	-0.00
Remote_Work	0.01	-0.00	-0.00	1.00	0.00	0.00	0.00	0.00
Publications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

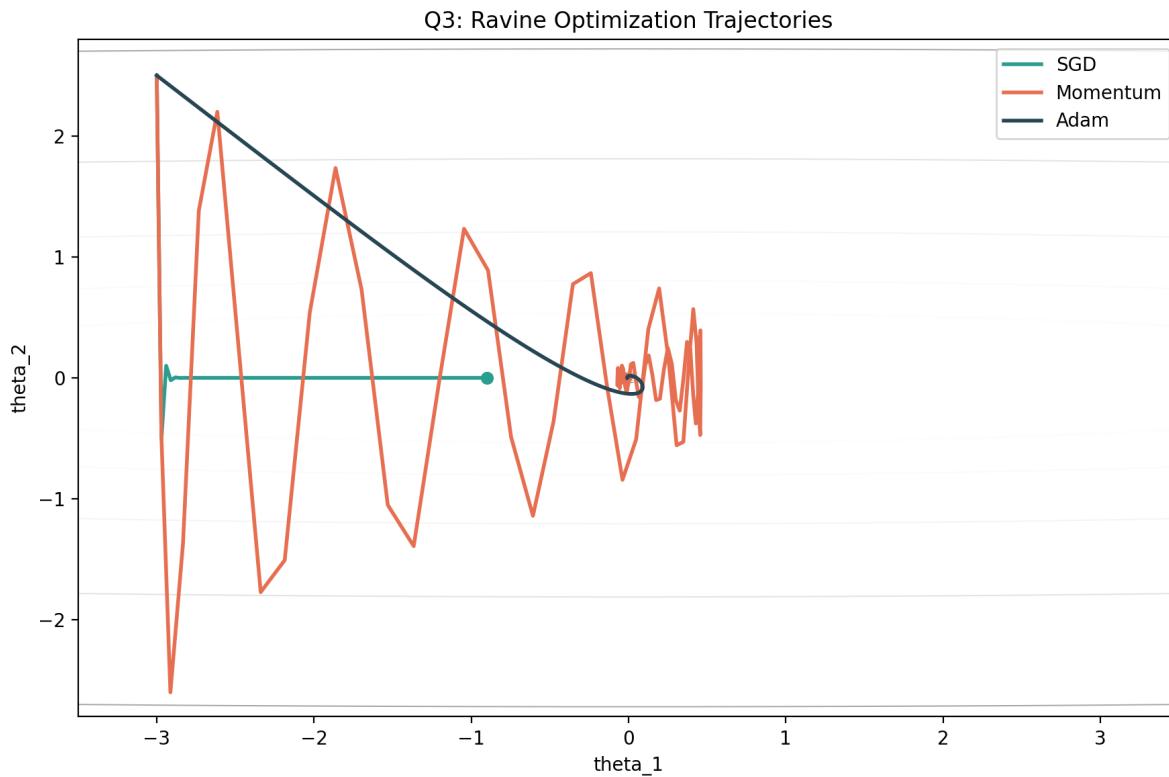
حدودیت/تهدید: تفاوت‌ها می‌توانند ناشی از محدودیت‌های سیاستی باشند نه توان واقعی افراد.

#### 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.

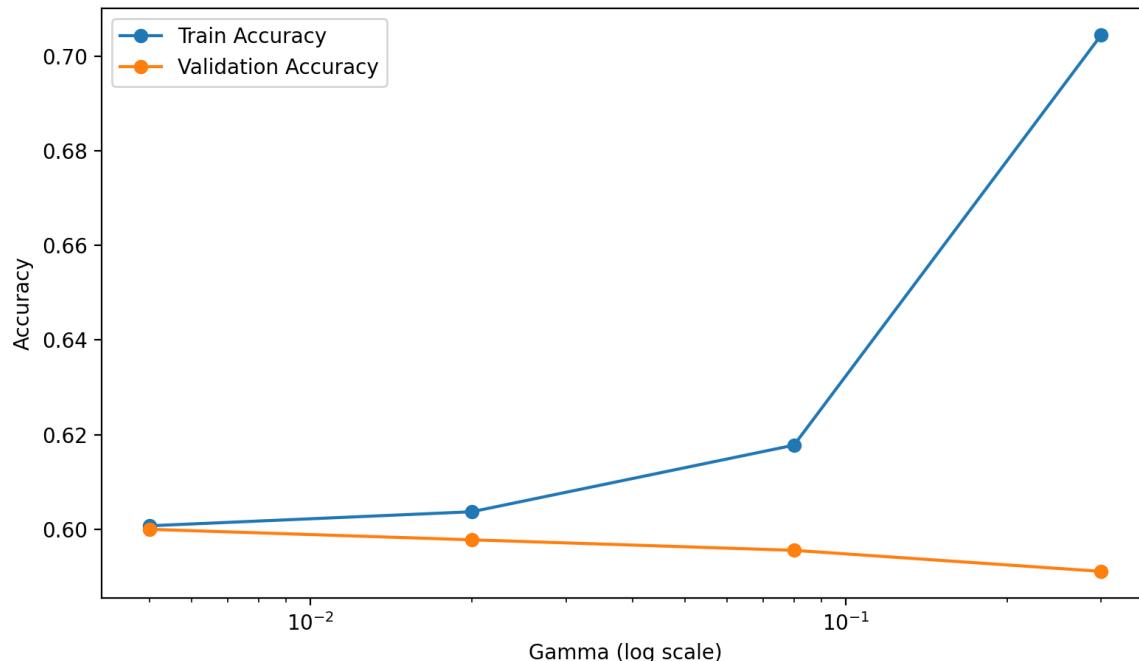
**تفسیر:** Momentum/Adam نوسان عرضی را کاهش می‌دهند و همگرایی سریع‌تری دارند.

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

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

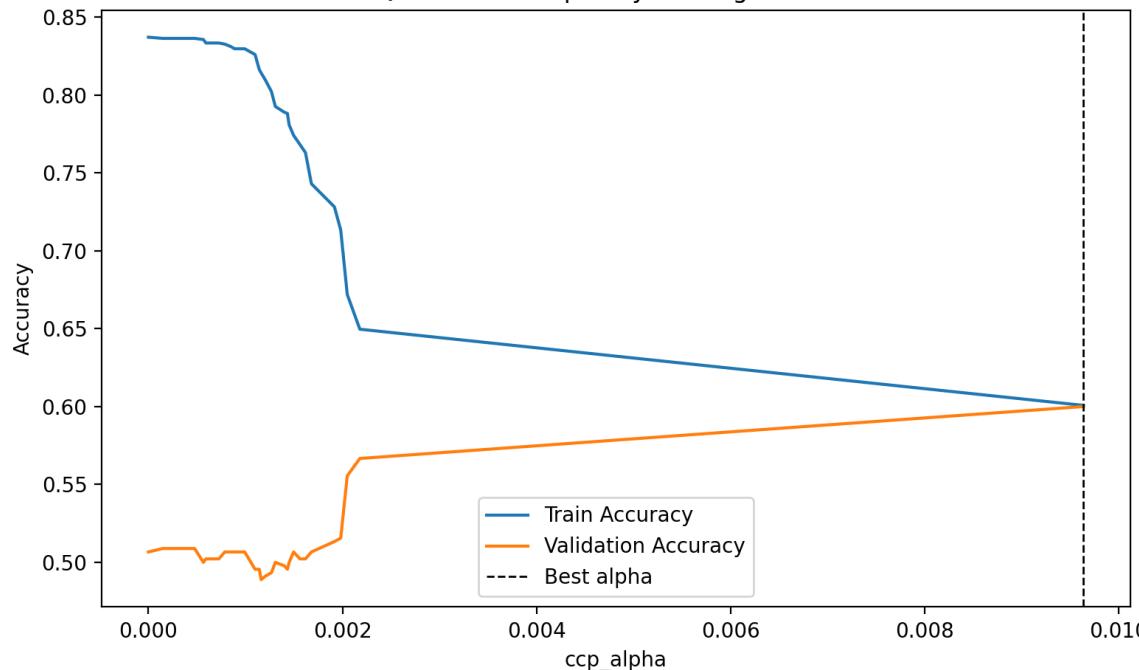
**Q4:** مدل غیرخطی و کنترل پیچیدگی

Q4A: SVM RBF Gamma vs Performance

شکل ۶: حساسیت دقت اعتبارسنجی نسبت به  $\gamma$  در SVM-RBF.

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

Q4B: Cost-Complexity Pruning Tradeoff

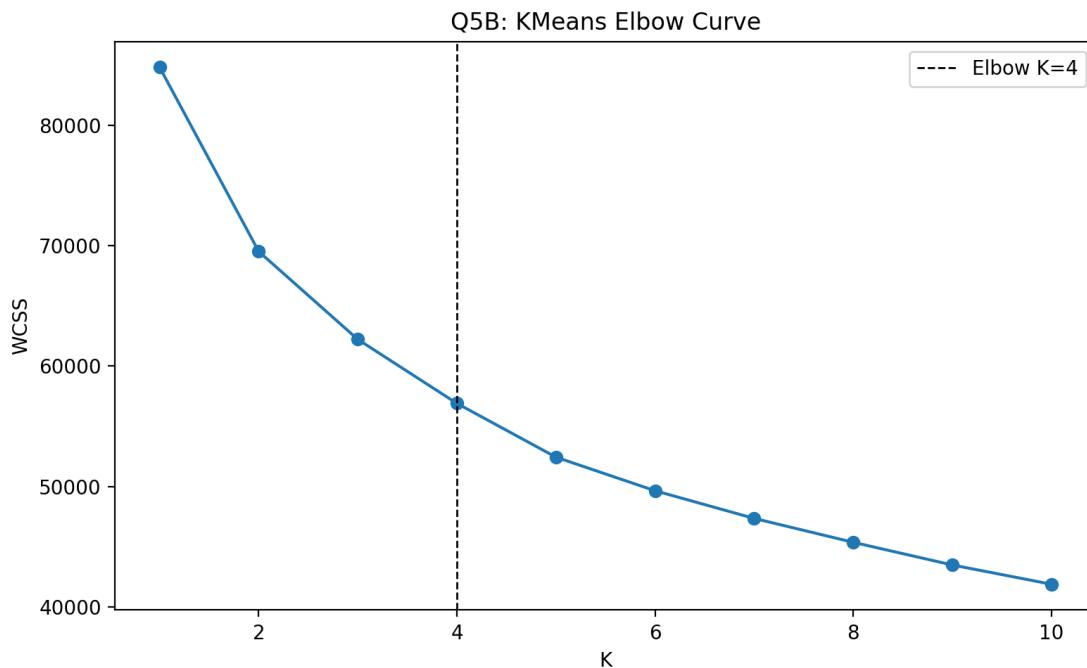


شکل ۷: مصالحه خط‌پیچیدگی در هرس Decision Tree

تفسیر:  $\alpha$  بزرگتر اندازه درخت را کم می‌کند و واریانس را پایین می‌آورد.  
اثر تصمیمی: انتخاب  $\alpha$  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

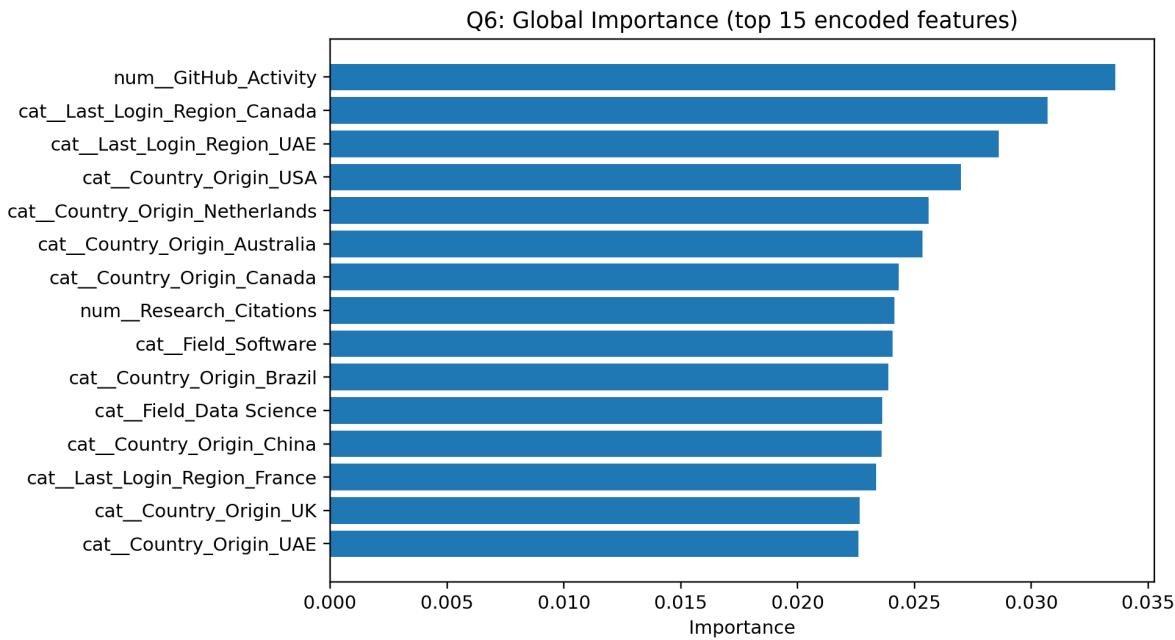


شکل ۹: توضیح محلی SHAP برای نمونه منتخب.

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.

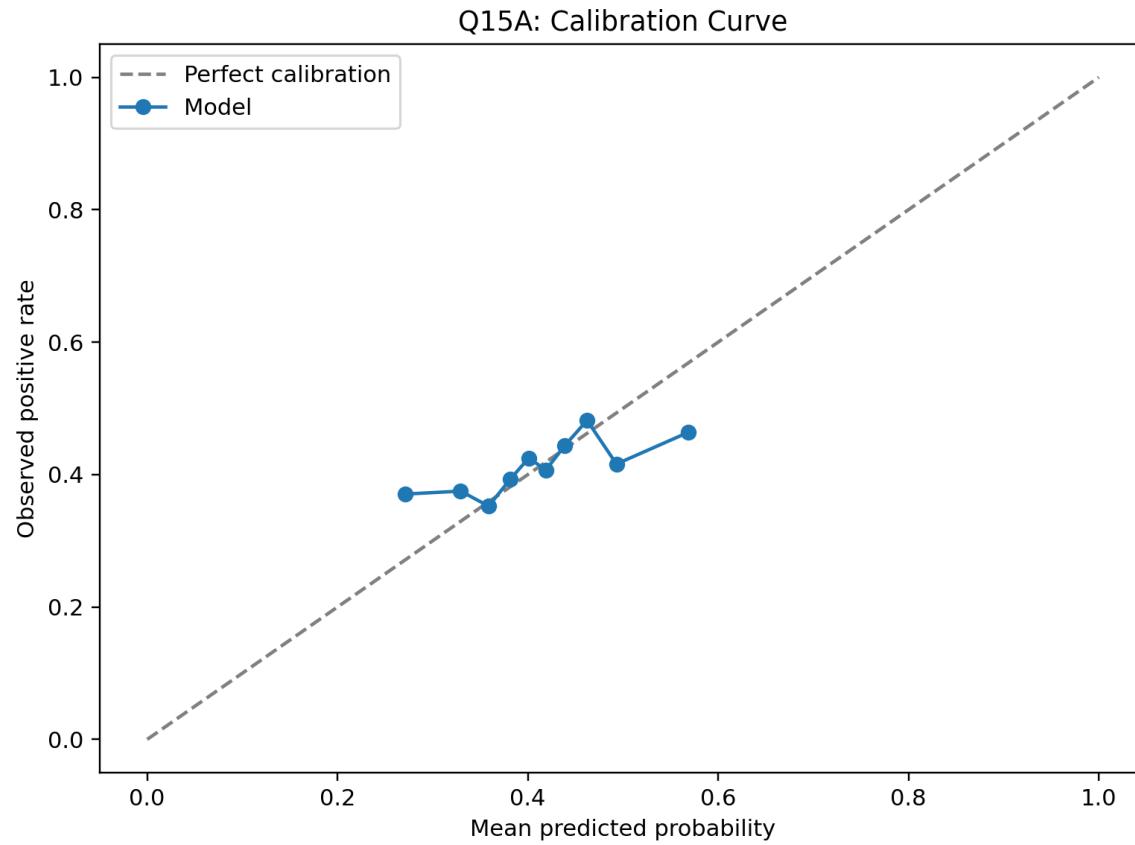
اثر تصمیمی: کنترل کیفیت داده برای ویژگی‌های پراثر، اولویت حاکمیتی می‌گیرد.

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

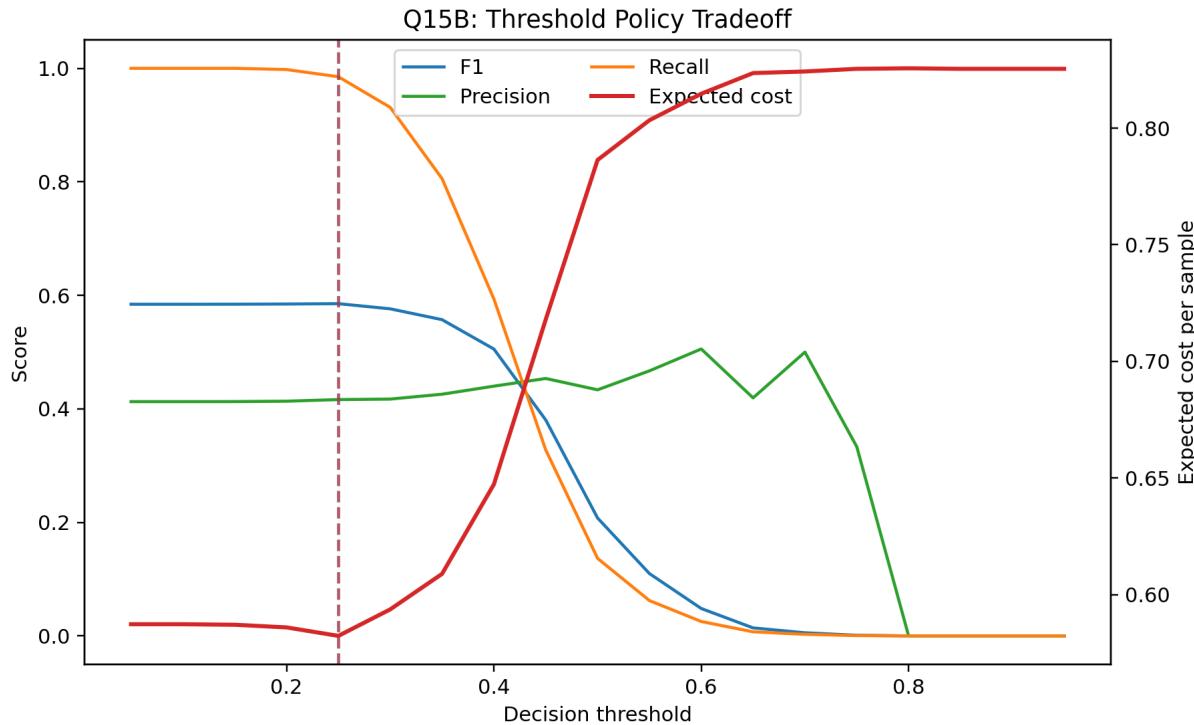


شکل ۱۱: منحنی قابلیت اطمینان کالیبراسیون احتمال.

**تفسیر:** کالیبراسیون نشان می‌دهد احتمال پیش‌بینی شده تا چه حد با فراوانی واقعی هم‌استناست.

**اثر تصمیمی:** تصمیم‌های حساس با احتمال کالیبرهشده گرفته می‌شوند، نه امتیاز خام.

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

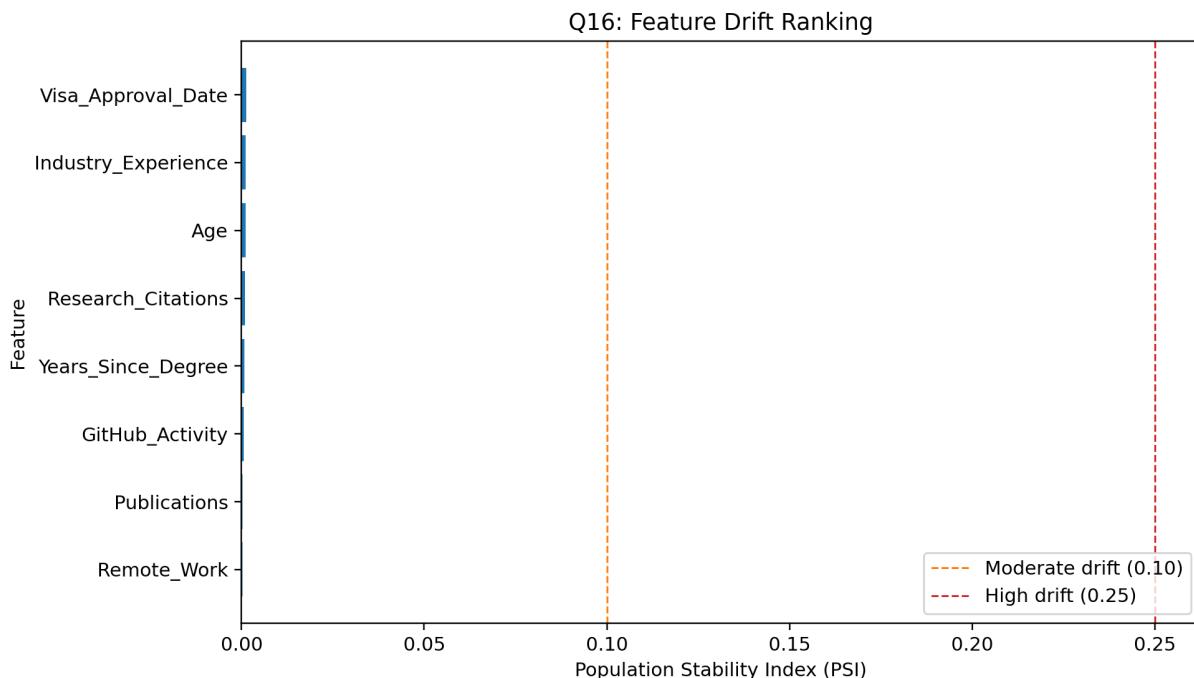


شکل ۱۲: مصالحه آستانه بین دقت، بازخوانی، F1 و هزینه مورد انتظار.

**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

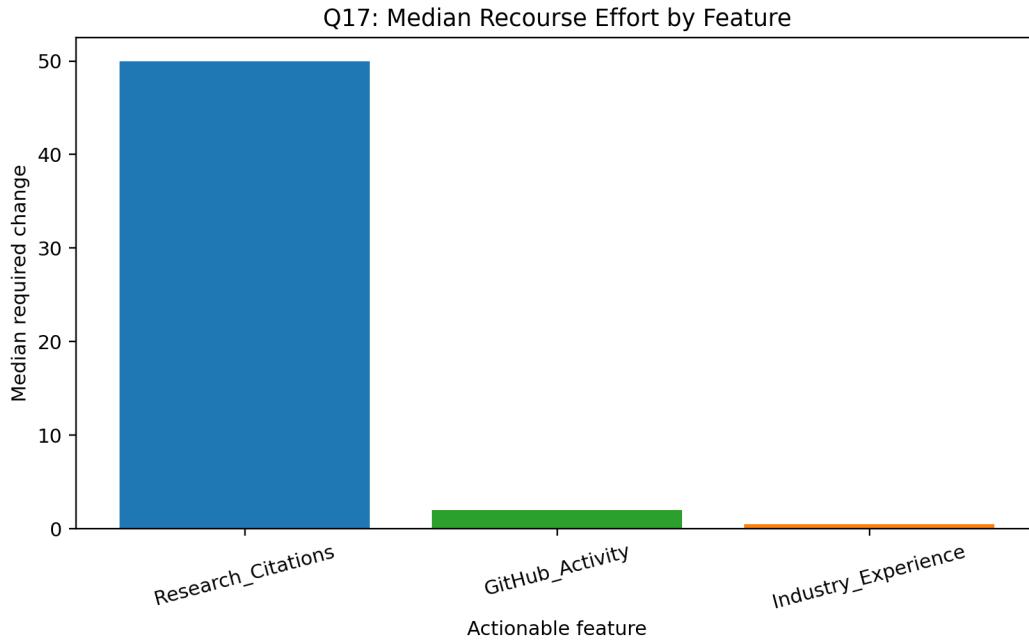
ویژگی با بیشترین درفت:  $\text{Visa\_Approval\_Date} = \text{PSI} + 0.130$



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

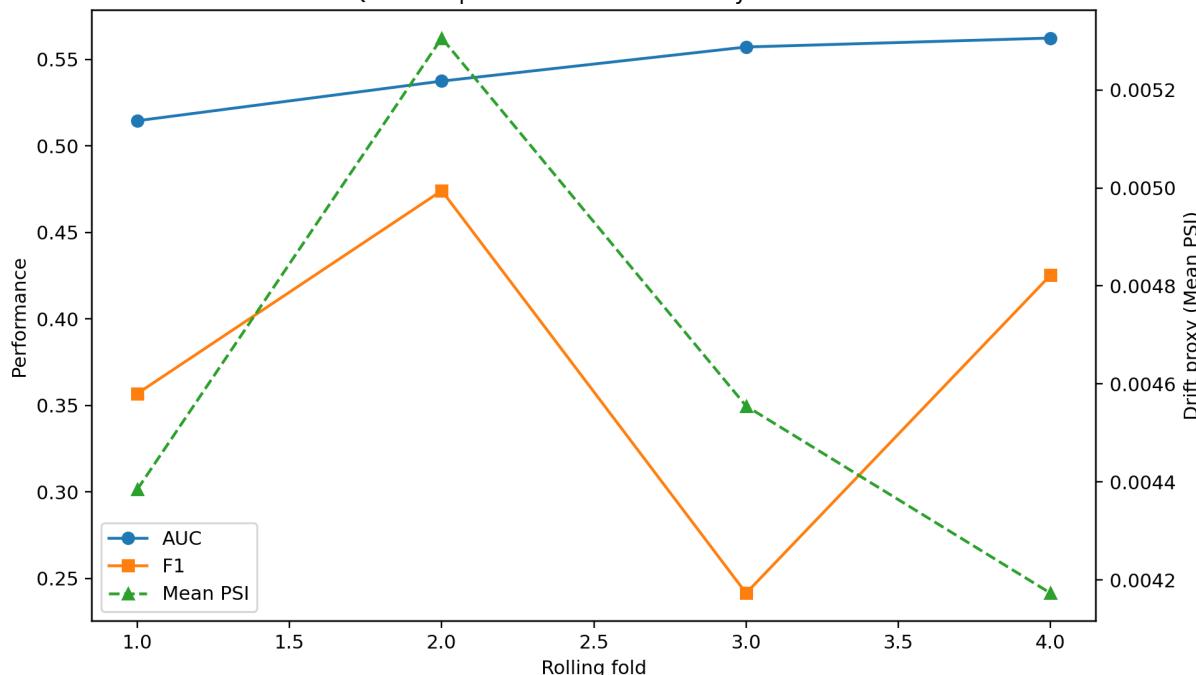


شکل ۱۴: میانه تغییر لازم برای عبور از مرز تصمیم در هر ویژگی قابل اقدام.

**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: .۵۴۲۸۰، افت AUC: .۰۴۷۸۰

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.

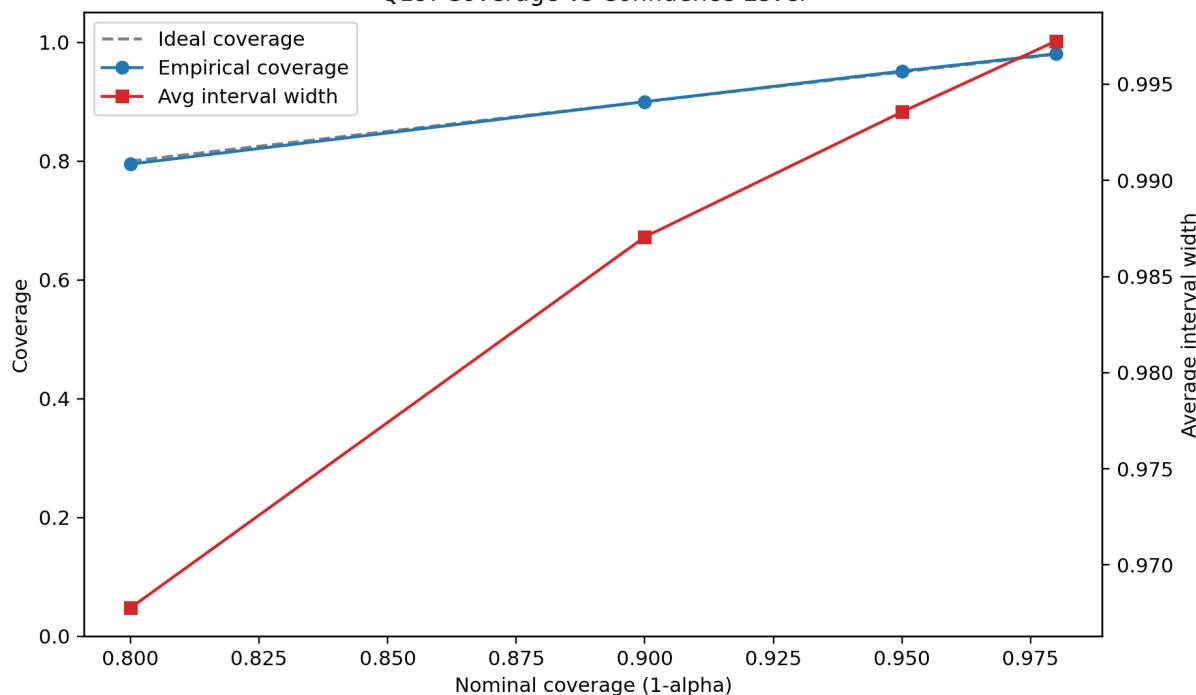
**اثر نصیبی:** قبل از استقرار، سنجش زمانی اجباری است و صرف random split کافی نیست.

**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.

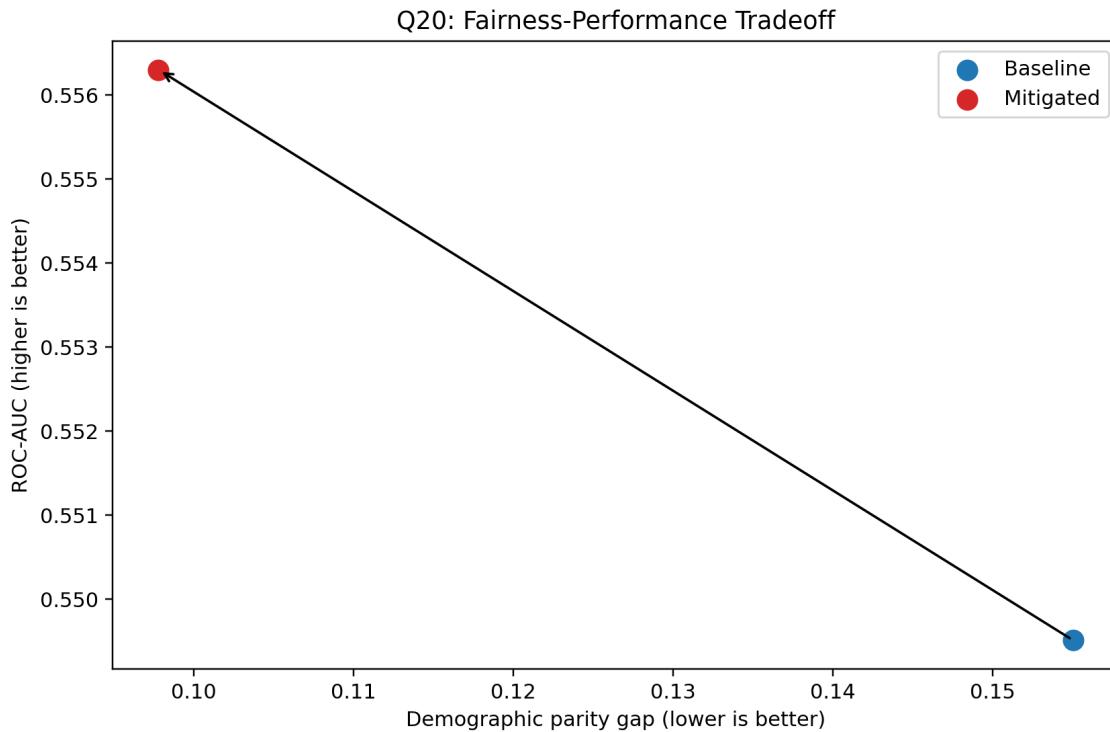
**تفسیر:** بازده‌های conformal توازن بین اطمینان پیش‌بینی و عرض بازه را روشن می‌کنند.

**اثر تصمیمی:** موارد کم اطمینان به مسیر بررسی انسانی ارجاع می‌شوند.

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

#### Q20: مداخله عدالت الگوریتمی

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.

#### بازتولیدپذیری و خروجی‌ها

اجرای کامل با دستور زیر انجام می‌شود:

`python code/scripts/full_solution_pipeline.py -profile {fast,balanced,heavy}`

خروجی‌ها در `code/figures` و `code/solutions` ذخیره می‌شوند، شامل `run_summary.json` نسخه ۲، خروجی‌های Q18-Q20 و فایل‌های `latex_metrics` برای گزارش خودکار.

#### جمع‌بندی

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