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
بسته کپستون کارشناسی‌ارشد – بهار ۱۴۰۴

چکیده

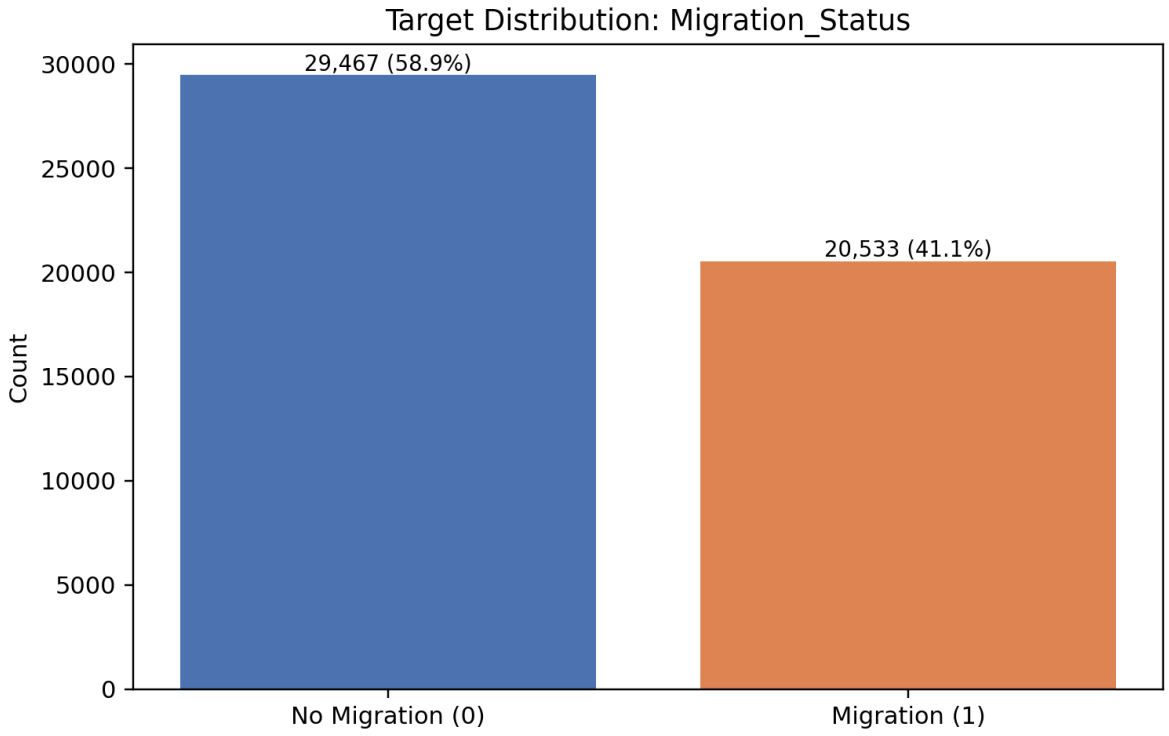
این گزارش نسخه کامل و قابل داوری پروژه نهایی درس علم داده است. پیاده‌سازی حاضر کل مسیر مهندسی داده تا تحلیل‌های تولیدی را پوشش می‌دهد: کنترل نشت، مدل‌سازی خطی و غیرخطی، بهینه‌سازی، یادگیری بدون نظارت، تبیین‌پذیری با SHAP، و سه افزونه پیشرفته Q18-Q20 برای پایداری زمانی، عدم‌قطعیت و عدالت الگوریتمی.

تعریف مسئله و دامنه پروژه

هدف، پیش‌بینی Migration_Status برای ۵۰ هزار متخصص فناوری است. خروجی نهایی صرفا 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

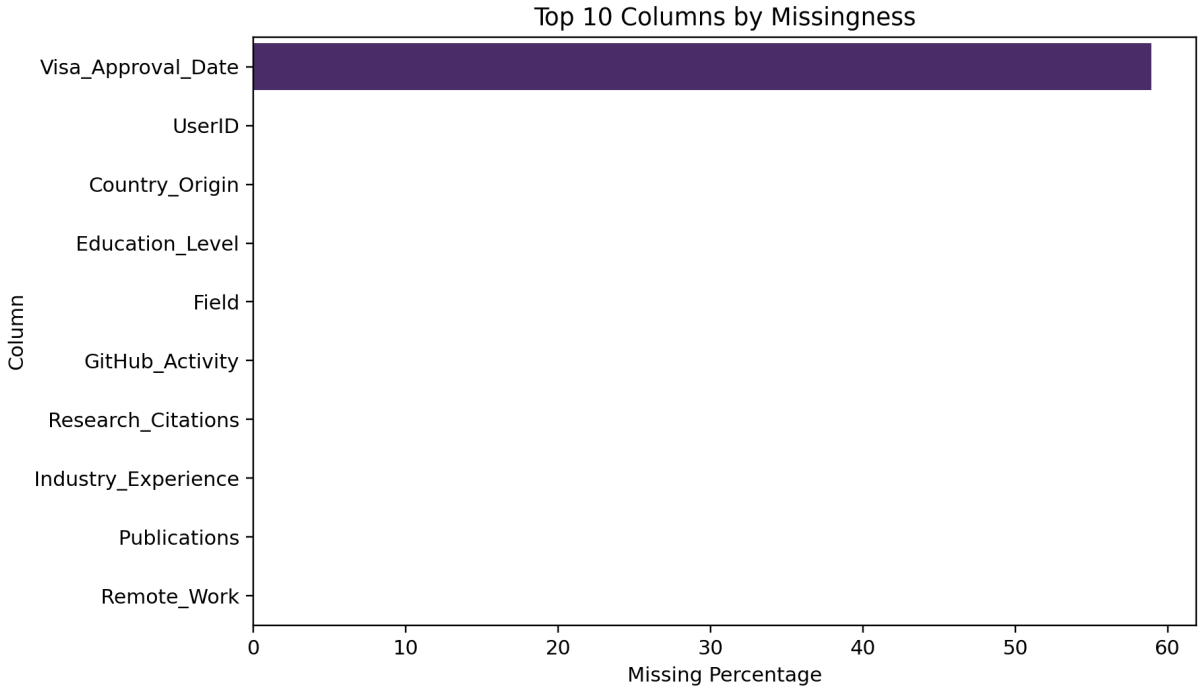
توازن کلاس هدف



شکل ۱: توزیع کلاس‌های متغیر هدف Migration_Status.

تفسیر: توزیع کلاس‌ها نامتوازنِ خفیف است و ارزیابی باید فراتر از دقت خام باشد.
اثر تصمیمی: معیارهای AUC/F1 و تحلیل آستانه، معیار اصلی انتخاب مدل قرار می‌گیرند.
محدودیت/تهدید: نسبت کلاس‌ها ممکن است در زمان استقرار تغییر کند و ثابت فرض‌کردن آن پرریسک است.

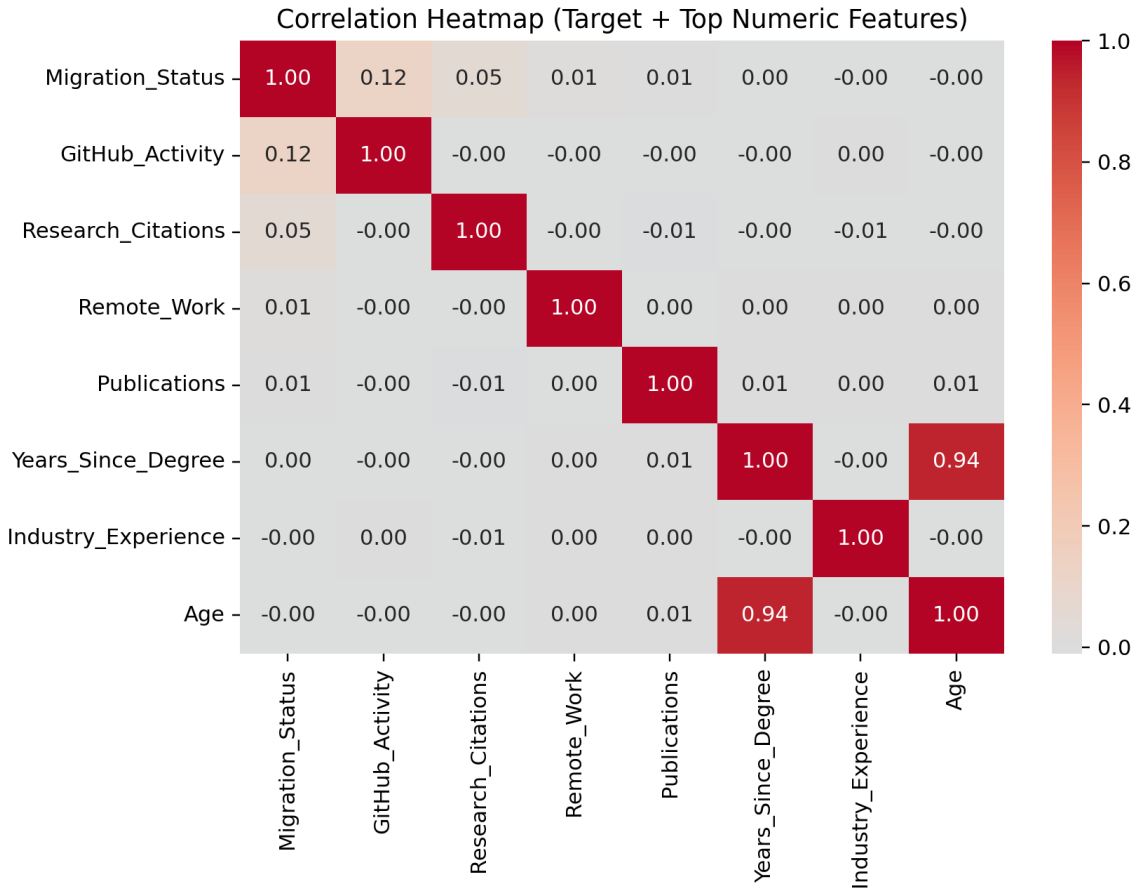
الگوی داده‌های گمشده



شکل ۲: ده ستون با بیشترین نرخ داده گمشده.

تفسیر: گمشدگی در ستون‌های فرایندی متمرکز است و بخشی از آن با نشت داده همپوشانی دارد.
اثر تصمیمی: ویژگی‌های پس‌رخداد، قبل از آموزش حذف می‌شوند یا به‌صورت کنترل‌شده وارد تحلیل می‌شوند.
محدودیت/تهدید: اگر گمشدگی تصادفی نباشد، خودِ الگوی گمشدگی می‌تواند منشأ بایاس شود.

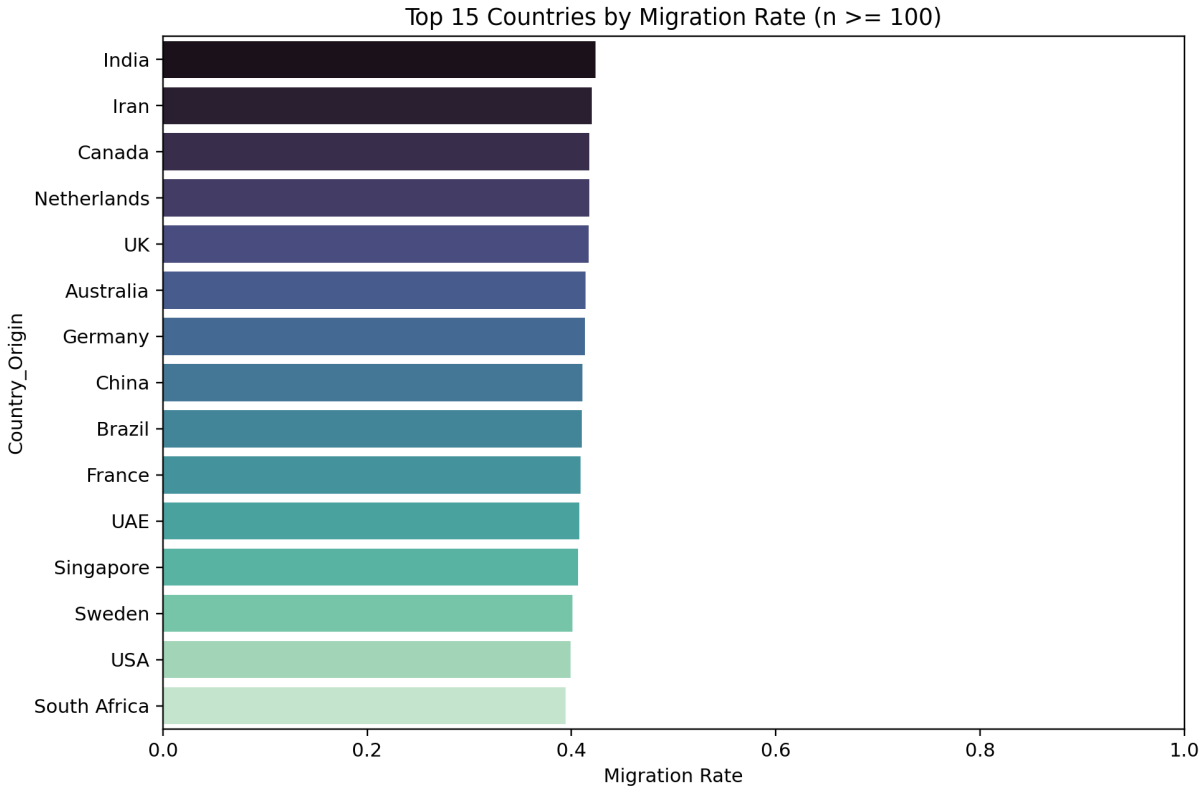
ساختار هم‌بستگی



شکل ۳: نقشه هم‌بستگی ویژگی‌های عددی کلیدی با هدف.

تفسیر: چند ویژگی مهم سیگنال دارند اما هیچ متغیری به‌تنهایی پاسخ کامل نمی‌دهد.
اثر تصمیمی: استفاده از مدل‌های چندمتغیره و غیرخطی به‌صورت هم‌زمان توجیه می‌شود.
محدودیت/تهدید: هم‌بستگی دلالت علّی 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.

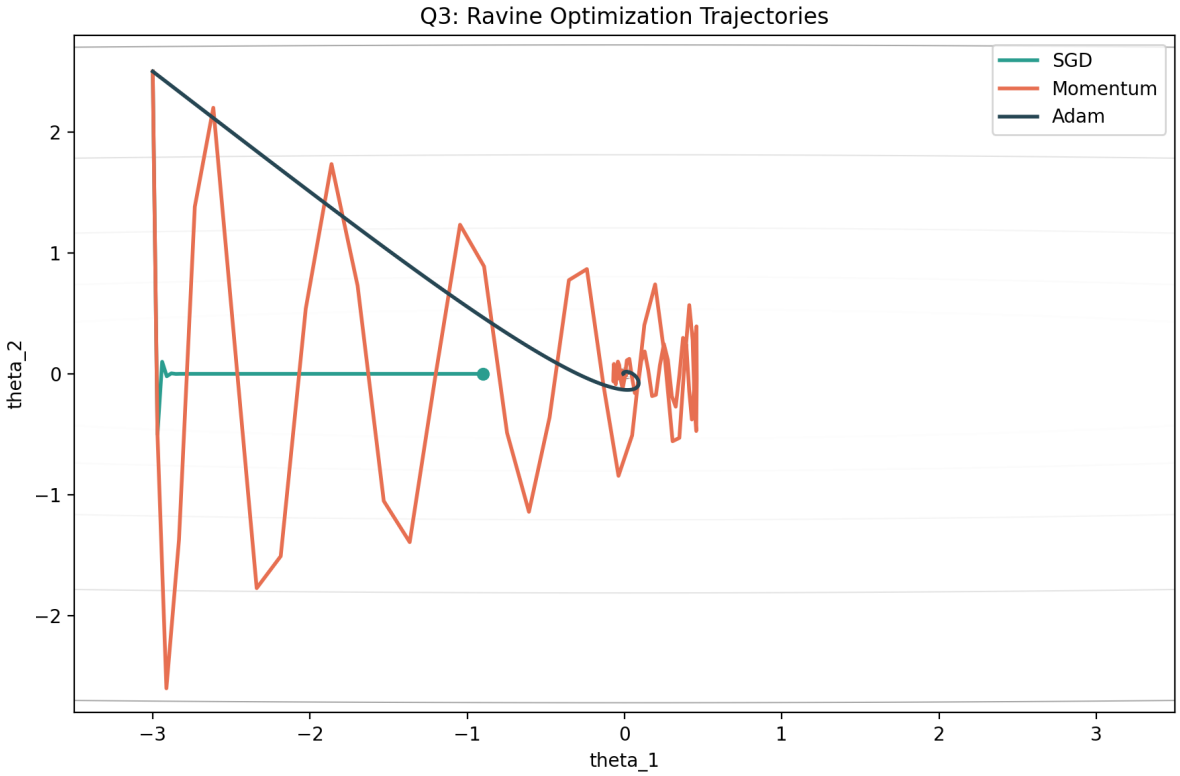
تفسیر: تفاوت گروهی بین کشورها معنادار است و بی‌توجهی به آن ارزیابی را ناقص می‌کند.
اثر تصمیمی: تحلیل عدالت قبل و بعد از مداخله الزامی است.
محدودیت/تهدید: تفاوت‌ها می‌توانند ناشی از محدودیت‌های سیاستی باشند نه توان واقعی افراد.

نتایج هسته اصلی (Q۱ تا Q۶)

Q۱: مهندسی داده و نشت

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

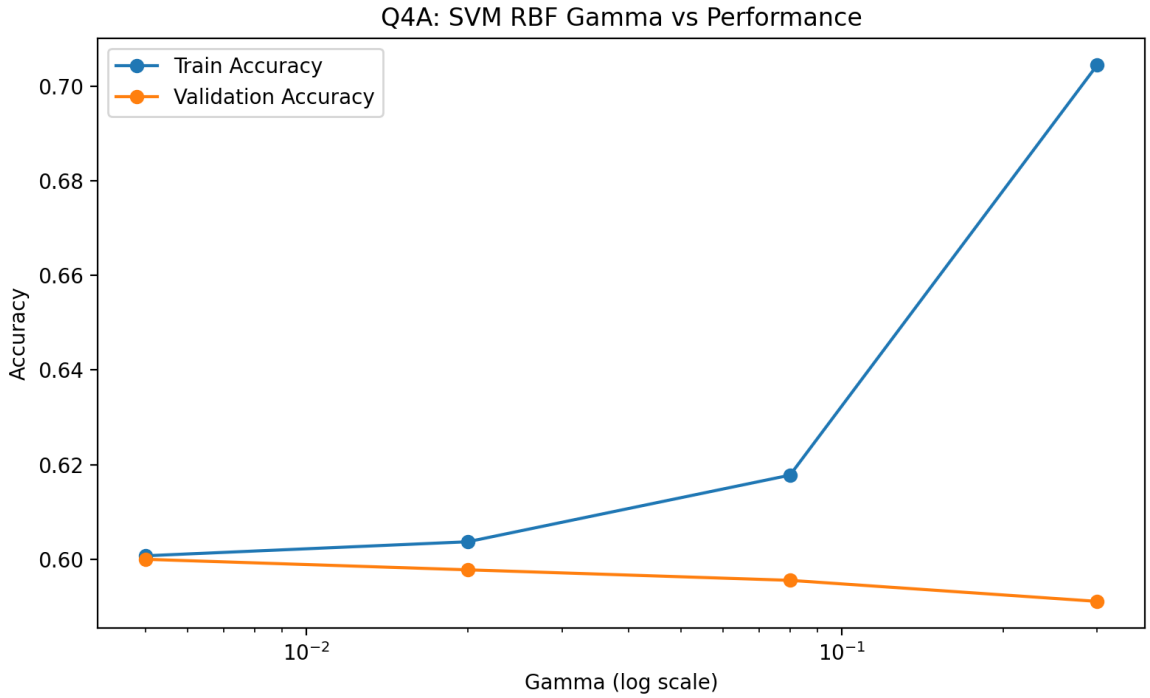
Q۳: تحلیل بهینه‌سازها



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

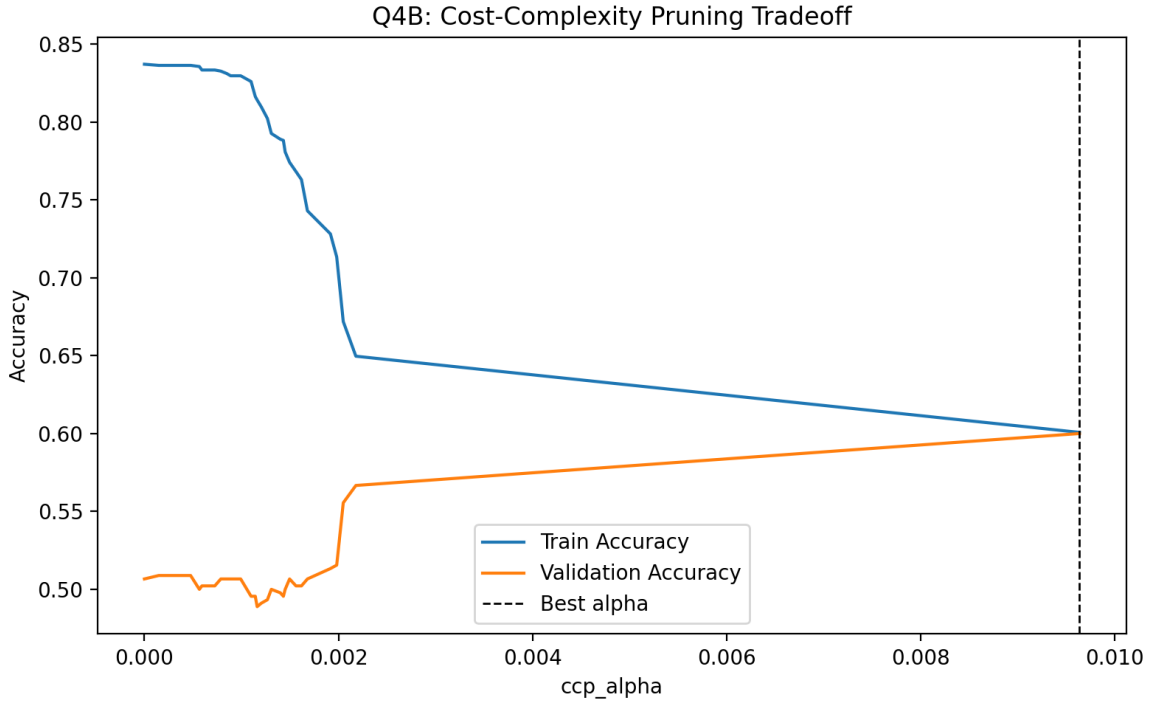
تفسیر: Momentum/Adam نوسان عرضی را کاهش می‌دهند و همگرایی سریع‌تری دارند.
اثر تصمیمی: برای توابع بدشرط، بهینه‌ساز شتاب‌دار انتخاب پیش‌فرض مناسب‌تری است.
محدودیت/تهدید: رفتار روی تابع آزمایشی، کل پیچیدگی اهداف غیرمحدب را پوشش نمی‌دهد.

Q۴: مدل غیرخطی و کنترل پیچیدگی



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

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

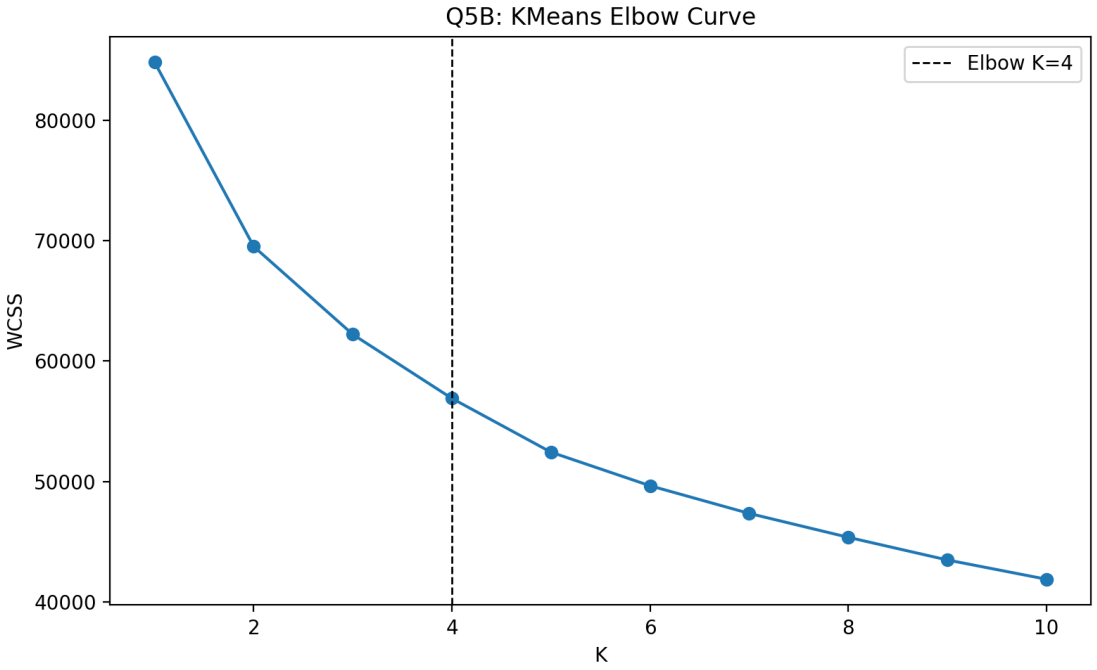


شکل ۷: مصالحه خطا-پیچیدگی در هرس Decision Tree.

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 آموزش.

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

Q۵: ساختار بدون نظارت

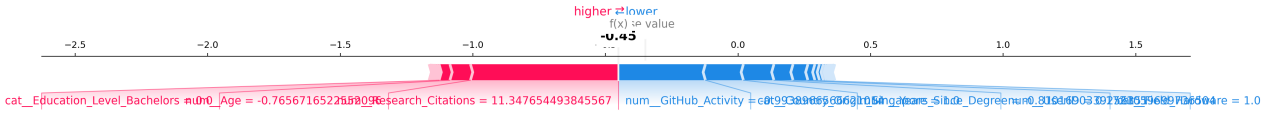


شکل ۸: منحنی آرنج WCSS برای انتخاب تعداد خوشه‌ها.

تفسیر: پس از یک K میانی، کاهش WCSS بازده نزولی پیدا می‌کند.
اثر تصمیمی: K به‌صورت heuristic انتخاب و سپس با تفسیرپذیری خوشه‌ها اعتبارسنجی می‌شود.
محدودیت/تهدید: در منحنی‌های کم‌انحنا، محل آرنج می‌تواند مبهم باشد.

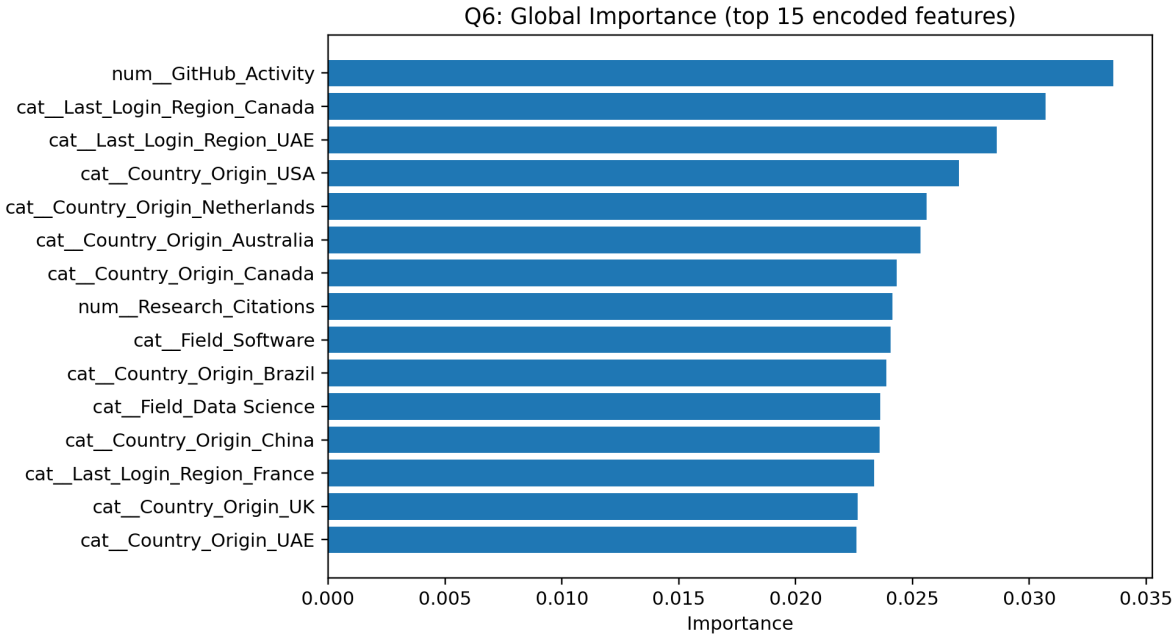
Q۶: مدل کپستون و SHAP

پروفایل اجرا: **balanced**.
مدل کپستون: **XGBoost**.
AUC: **۰.۵۴۹۵**، Accuracy: **۰.۵۸۳۵**، F1: **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 متمرکز می‌شود.
محدودیت/تهدید: SHAP رفتار مدل را توضیح می‌دهد، نه رابطه علّی دنیای واقعی را.



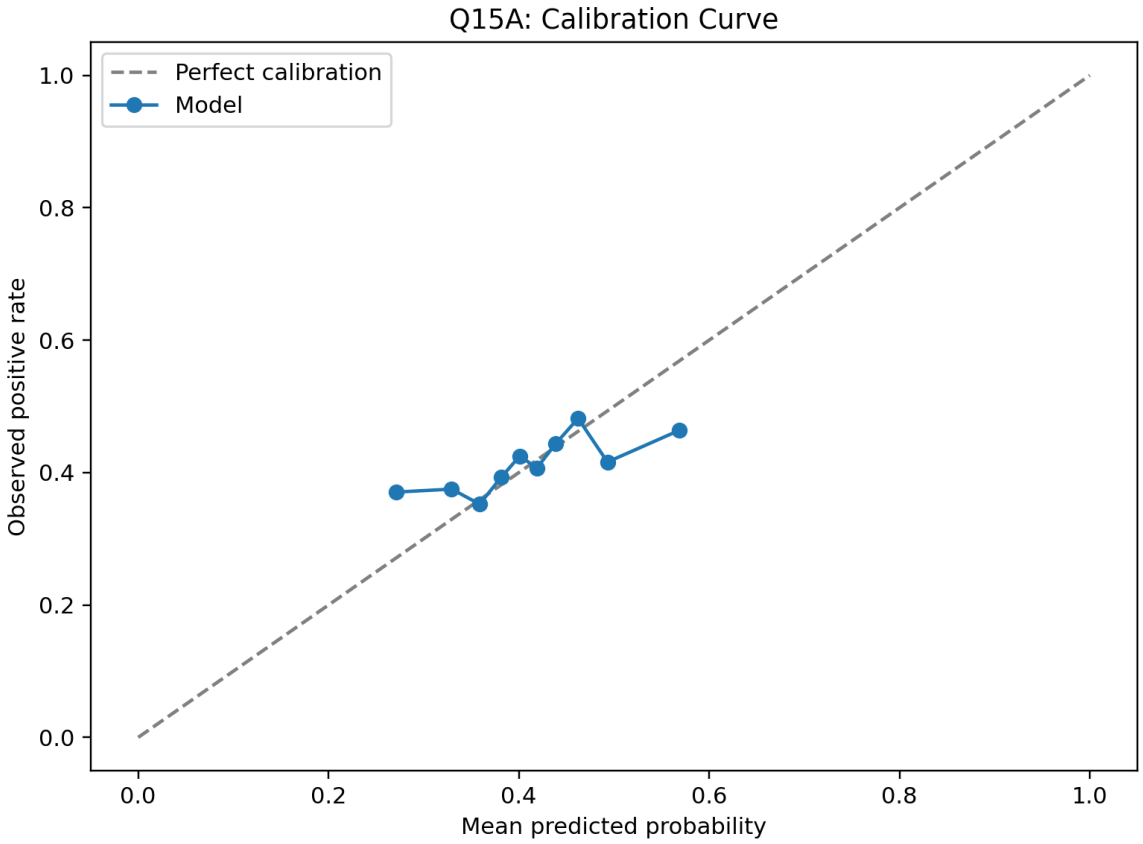
شکل ۱۰: نمای سراسری اهمیت ویژگی‌ها در مدل کپستون.

تفسیر: ویژگی‌های مرتبط با فعالیت فنی و استناد، وزن بیشتری در پیش‌بینی دارند.
اثر تصمیمی: کنترل کیفیت داده برای ویژگی‌های پراثر، اولویت حاکمیتی می‌گیرد.
محدودیت/تهدید: اهمیت سراسری، ناهمگنی اثر در همه زیرگروه‌ها را نشان نمی‌دهد.

بلوک پیشرفته تولیدی (Q۱۵ تا Q۲۰)

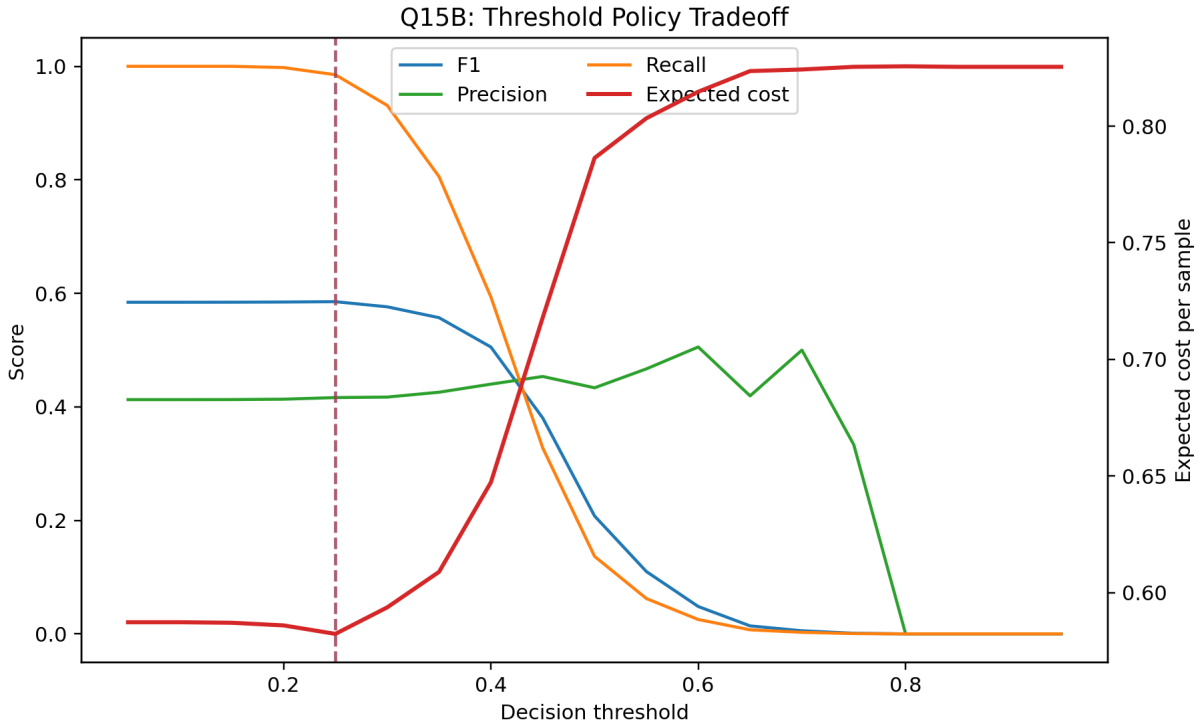
Q۱۵: کالیبراسیون و سیاست آستانه

Brier: ۰.۲۴۳۶، ECE: ۰.۰۳۲۷، آستانه بهینه F1: ۰.۲۵۰۰.



شکل ۱۱: منحنی قابلیت اطمینان کالیبراسیون احتمال.

تفسیر: کالیبراسیون نشان می‌دهد احتمال پیش‌بینی‌شده تا چه حد با فراوانی واقعی هم‌راستاست.
اثر تصمیمی: تصمیم‌های حساس با احتمال کالیبره‌شده گرفته می‌شوند، نه امتیاز خام.
محدودیت/تهدید: کالیبراسیون در گذر زمان ممکن است فرسوده شود.

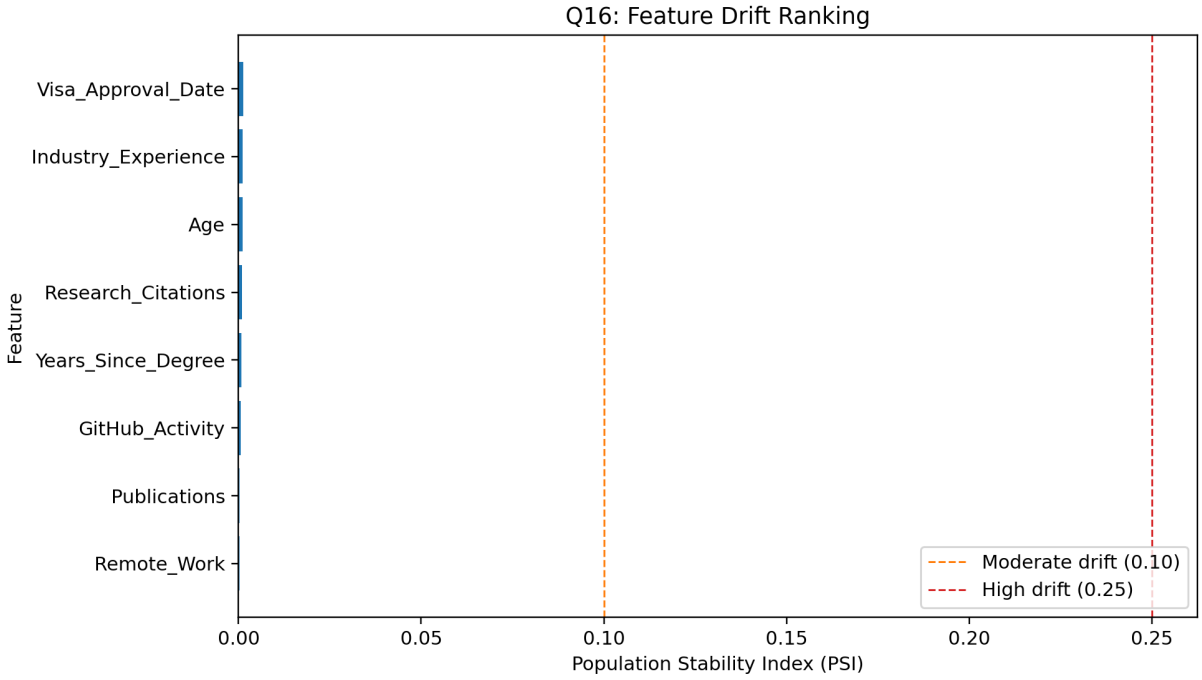


شکل ۱۲: مصالحه آستانه بین دقت، بازخوانی، F1 و هزینه مورد انتظار.

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

Q۱۶: پایش درفت

ویژگی با بیشترین درفت: **Visa_Approval_Date** با PSI=۰.۰۰۱۳.

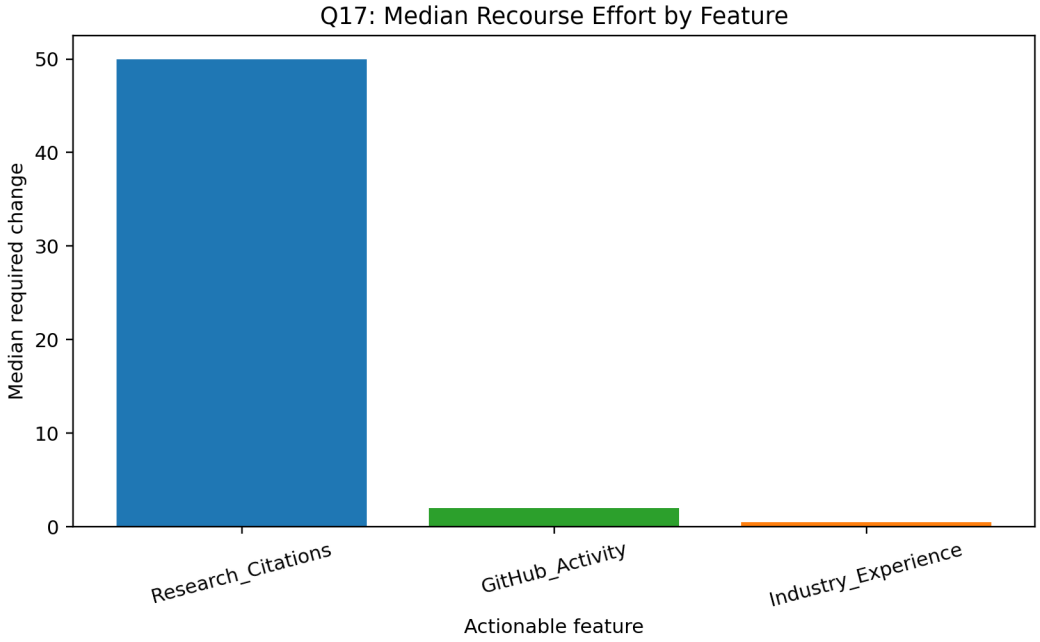


شکل ۱۳: رتبه‌بندی درفت ویژگی‌ها بر اساس PSI.

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

Q۱۷: ریکورس مقابله‌ای

نرخ موفقیت ریکورس: **۱.۰۰۰۰**.

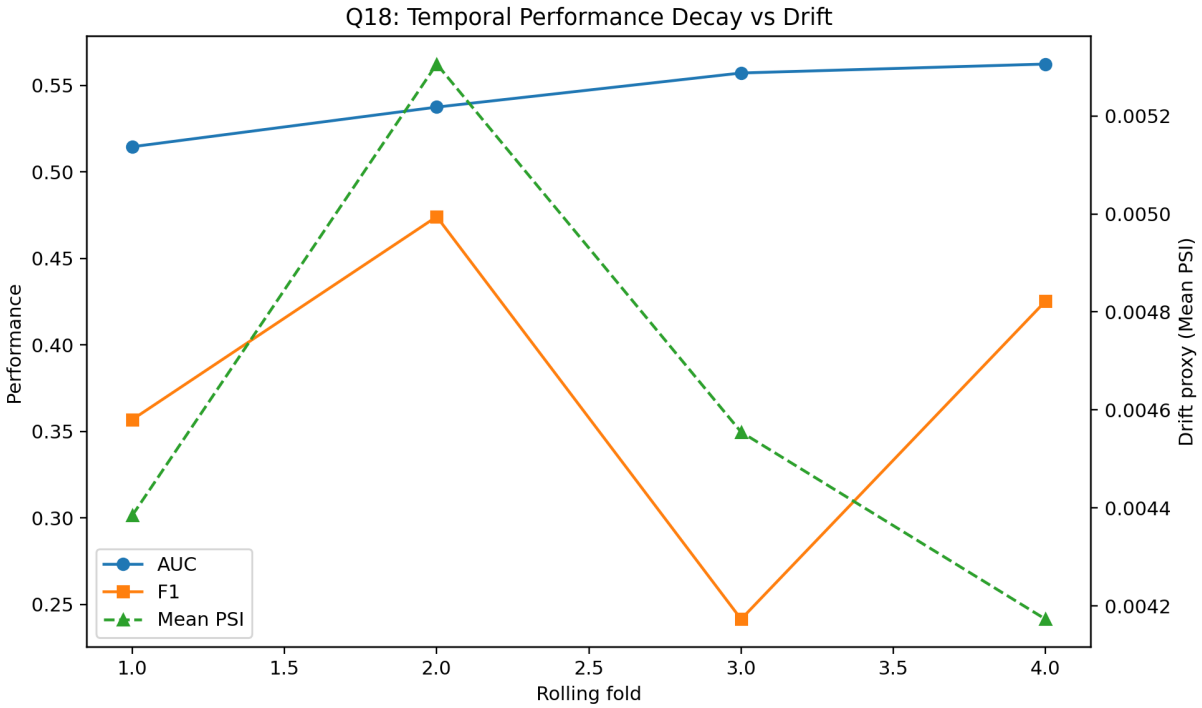


شکل ۱۴: میانه تغییر لازم برای عبور از مرز تصمیم در هر ویژگی قابل اقدام.

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

Q۱۸: اعتبارسنجی زمانی و افت عملکرد

میانگین AUC زمانی: **۰.۵۴۲۸**، افت AUC: **۰.۰۴۷۸**.

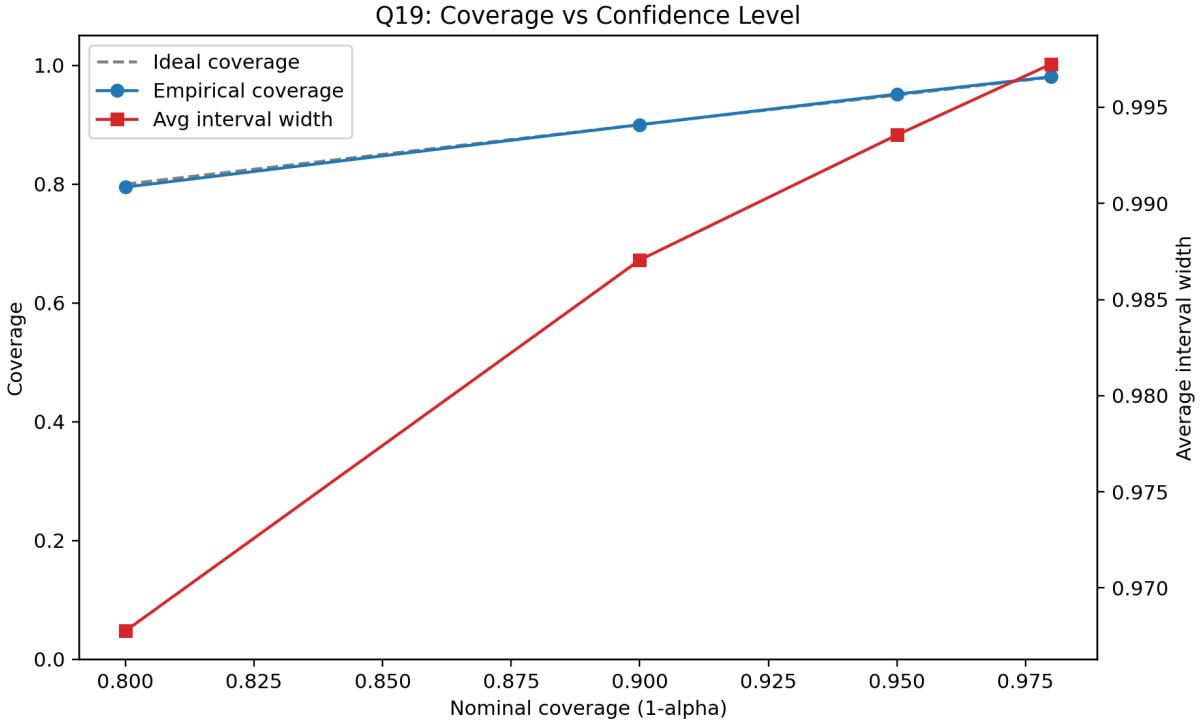


شکل ۱۵: افت عملکرد در پنجره‌های زمانی غلطان در کنار شاخص درفت.

تفسیر: عملکرد بین هایfold متوالی تغییر می‌کند و با شدت درفت قابل مقایسه است.
اثر تصمیمی: قبل از استقرار، سنجش زمانی اجباری است و صرف random split کافی نیست.
محدودیت/تهدید: در نبود ستون زمانی معتبر، fallback باید صریح گزارش شود.

Q۱۹: کمی‌سازی عدم‌قطعیت

Coverage@90: ۰.۹۰۰۰، بیشینه کم‌پوششی: ۰.۰۰۵۰.

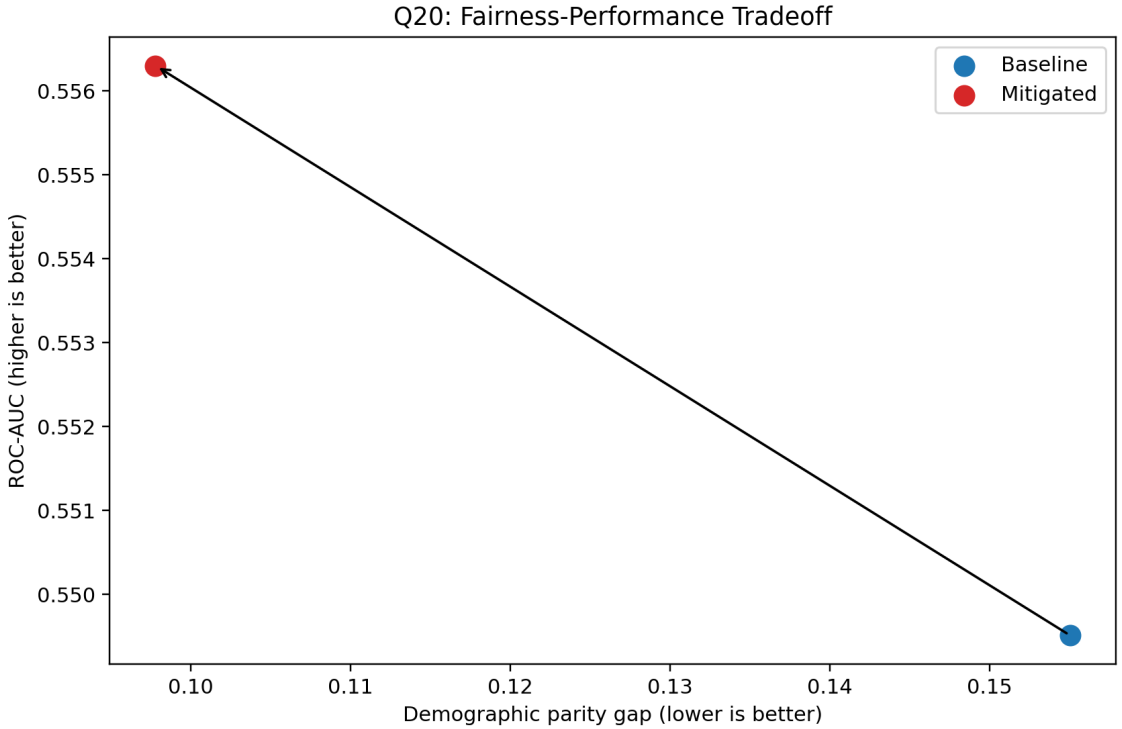


شکل ۱۶: پوشش اسمی در برابر پوشش تجربی و پهنای بازه عدم‌قطعیت.

تفسیر: بازه‌های conformal توازن بین اطمینان پیش‌بینی و عرض بازه را روشن می‌کنند.
اثر تصمیمی: موارد کم‌اطمینان به مسیر بررسی انسانی ارجاع می‌شوند.
محدودیت/تهدید: تضمین‌های پوشش در صورت جابه‌جایی توزیع تضعیف می‌شوند.

Q۲۰: مداخله عدالت الگوریتمی

شکاف DP پایه: ۰.۱۵۵۰، شکاف DP پس از مداخله: ۰.۰۹۷۸، نتیجه قید سیاستی: **true**.



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

تفسیر: بازوزن‌دهی، نقطه عملکردی مدل را روی صفحه عدالت-کارایی جابه‌جا می‌کند.
اثر تصمیمی: استقرار منوط به عبور همزمان از قید افت عملکرد و بهبود عدالت است.
محدودیت/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/solutions و code/figures ذخیره می‌شوند، شامل run_summary.json نسخه ۲، خروجی‌های Q18-Q20 و فایل‌های latex_metrics برای گزارش خودکار.

ضمیمه A: تشریح کدهای پروژه

نقشه ماژول‌ها

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_uncertainty.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 متریک‌ها را به JSON/TEX تبدیل می‌کند تا گزارش‌ها hard-code نباشند.
generate_report_assets.py شکل‌های گزارش مدیریتی و report_stats.json را بازتولید می‌کند.

منطق مهندسی

پایپ‌لاین 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 انتها 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 بمانند.

ضمیمه B: شناسنامه کامل شکل‌ها

شکل‌های گزارش پایه داده

report_target_balance.png: هدف، بررسی توازن کلاس است. اگر عدم‌توازن زیاد باشد، دقت خام گمراه‌کننده می‌شود و باید از AUC/F1 استفاده شود.
report_missingness_top10.png: ستون‌های دارای گمشدگی بالا را نشان می‌دهد و ورودی تصمیم‌های حذف/جایگزینی است.
report_numeric_correlation.png: هم‌بستگی اولیه ویژگی‌ها با هدف را نمایش می‌دهد و برای اولویت‌بندی تحلیل چندمتغیره استفاده می‌شود.
report_country_migration_rate.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 روی 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 ccp_alpha بر tradeoff بایاس-واریانس 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: هم‌راستایی 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 بازآموزی.
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 پهنای بازه عدم‌قطعیت؛ 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 تکمیلی آموزشی

shap_force_plot.png :(legacy) خروجی نسخه baseline آموزشی در train_and_explain.py است و جایگزین تحلیل کپستون Q6 نیست.

جمع‌بندی

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