Model Testing Report

Experiment: model_testing

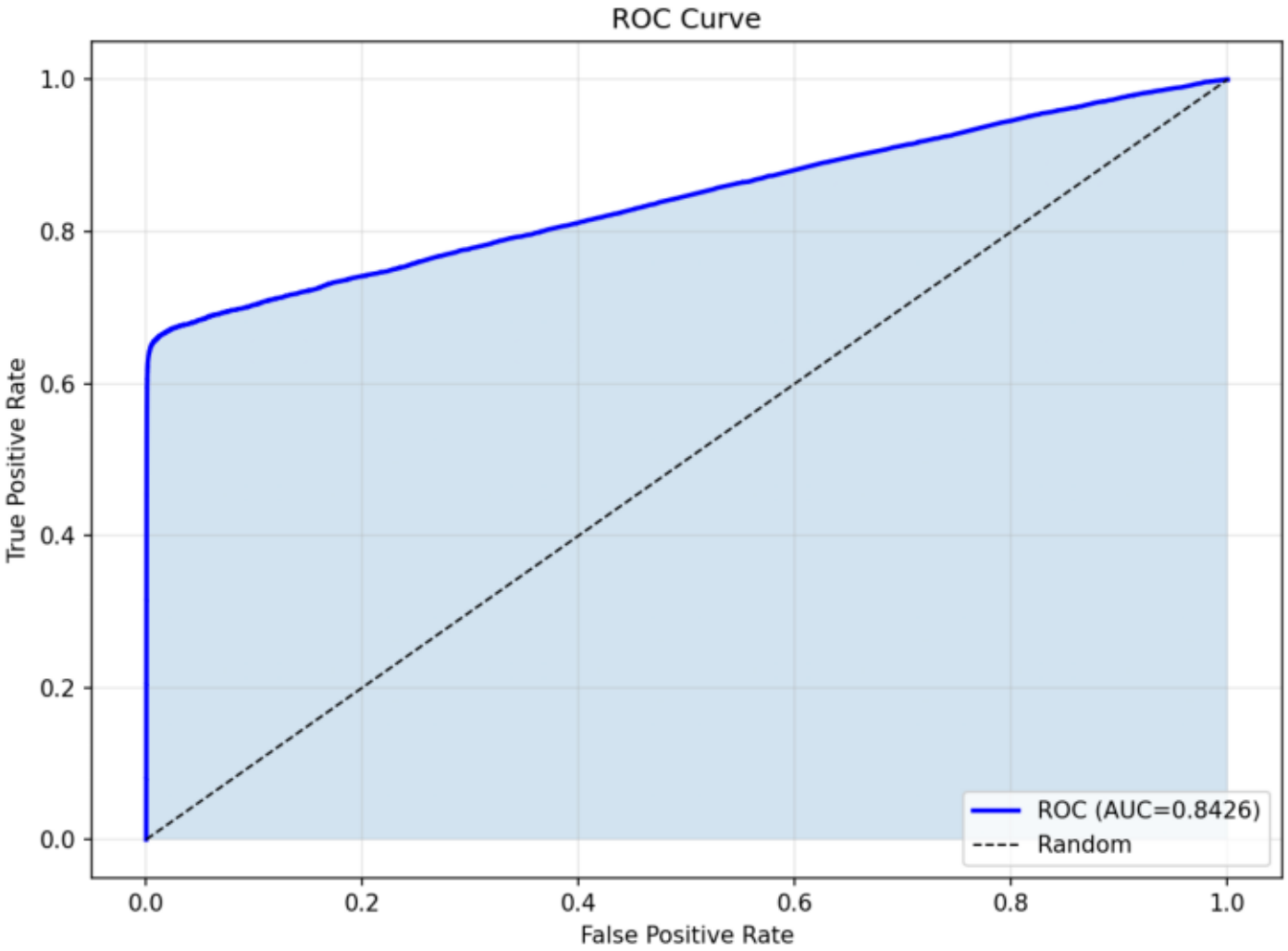
Generated: 2026-01-29 14:45:29

1) Model Effectiveness

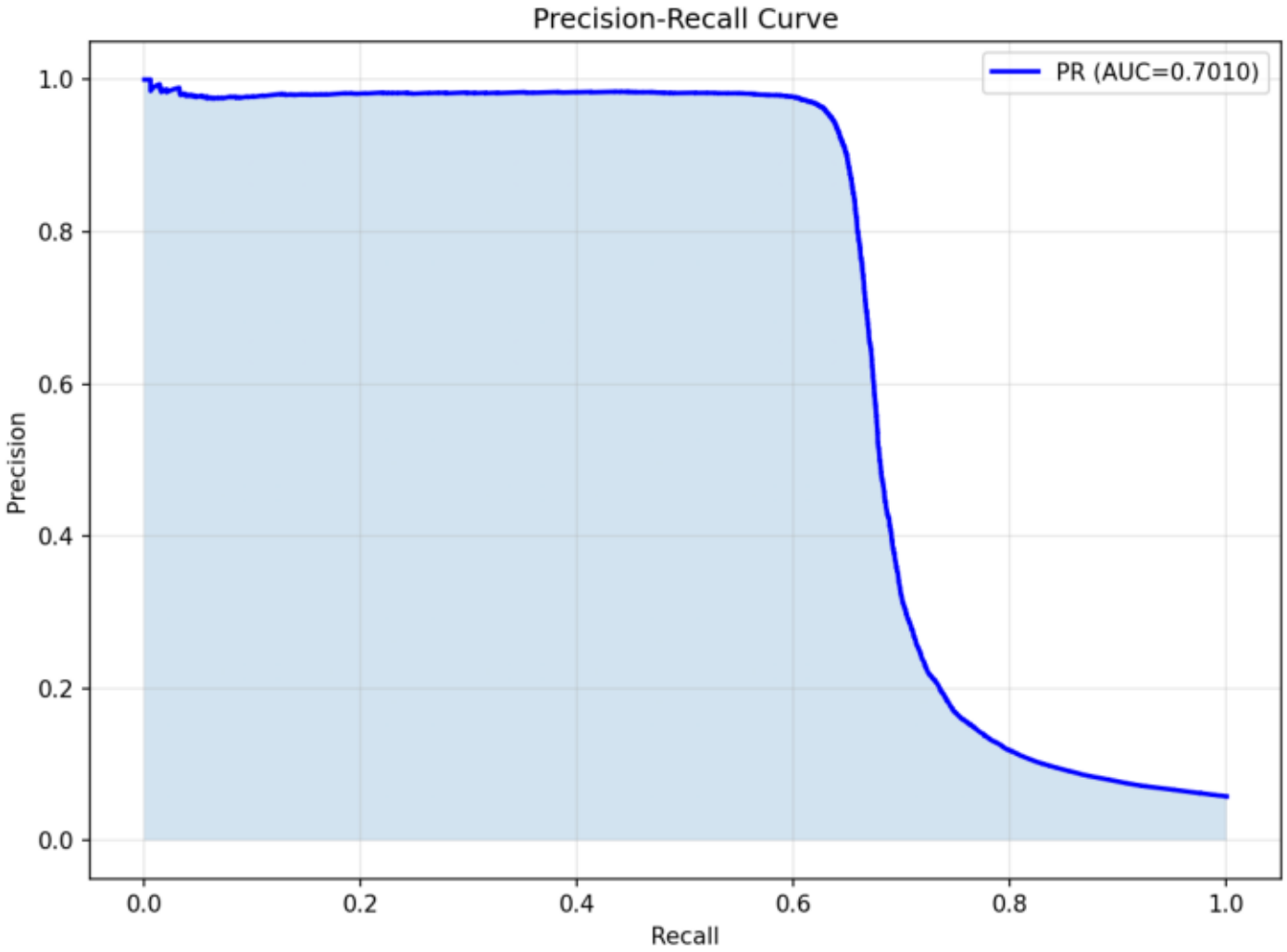
- auc_roc: 0.8426
- auc_pr: 0.7010
- precision: 0.9713
- recall: 0.6140
- f1: 0.7524
- ks_statistic: 0.6503
- ks_threshold: 0.0800
- confusion_matrix.TN: 188273
- confusion_matrix.FP: 209
- confusion_matrix.FN: 4446
- confusion_matrix.TP: 7072
- precision_at_k.10: 1.0000
- precision_at_k.50: 1.0000
- precision_at_k.100: 0.9900
- precision_at_k.200: 0.9850
- precision_at_k.500: 0.9800
- recall_at_k.10: 0.0009
- recall_at_k.50: 0.0043
- recall_at_k.100: 0.0086
- recall_at_k.200: 0.0171
- recall_at_k.500: 0.0425

1) Model Effectiveness - Explanations

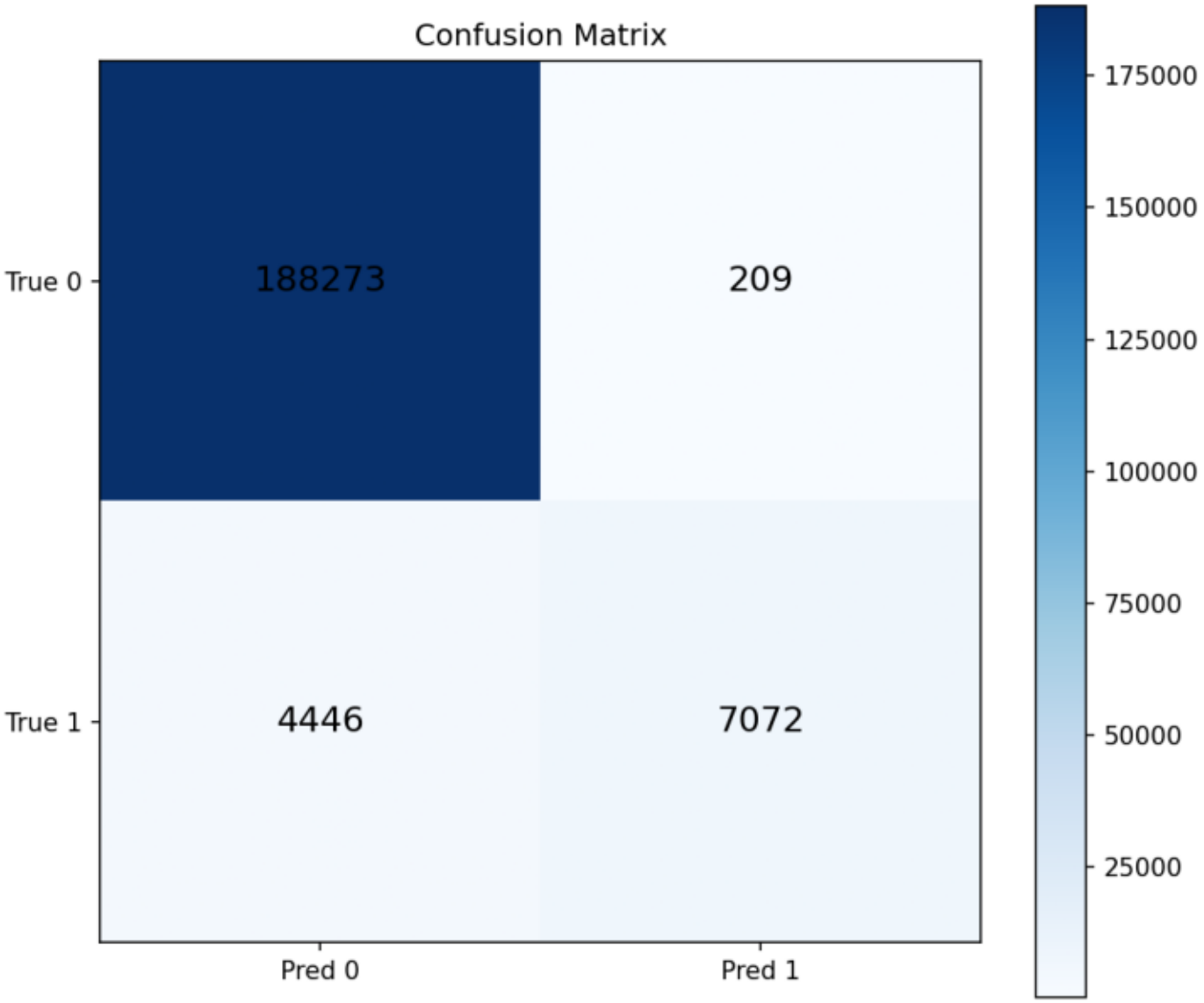
- Positive rate is 5.76%; AUC-ROC is 0.8426 (strong ranking).
- AUC-PR is 0.7010, above the base rate 5.76%, indicating meaningful lift.
- At threshold 0.50, precision=0.9713, recall=0.6140, F1=0.7524 (precision-heavy (fewer positives caught, fewer
- false alarms)).
- Confusion matrix: TP=7072, FP=209, FN=4446, TN=188273 (FPR=0.11%, FNR=38.60%).
- KS is 0.6503 at threshold 0.08, indicating very strong separation.
- Top-10: precision=1.0000, recall=0.0009 shows the quality of the highest-risk shortlist.
- Top-500: precision=0.9800, recall=0.0425 shows how much coverage you get with a larger queue.



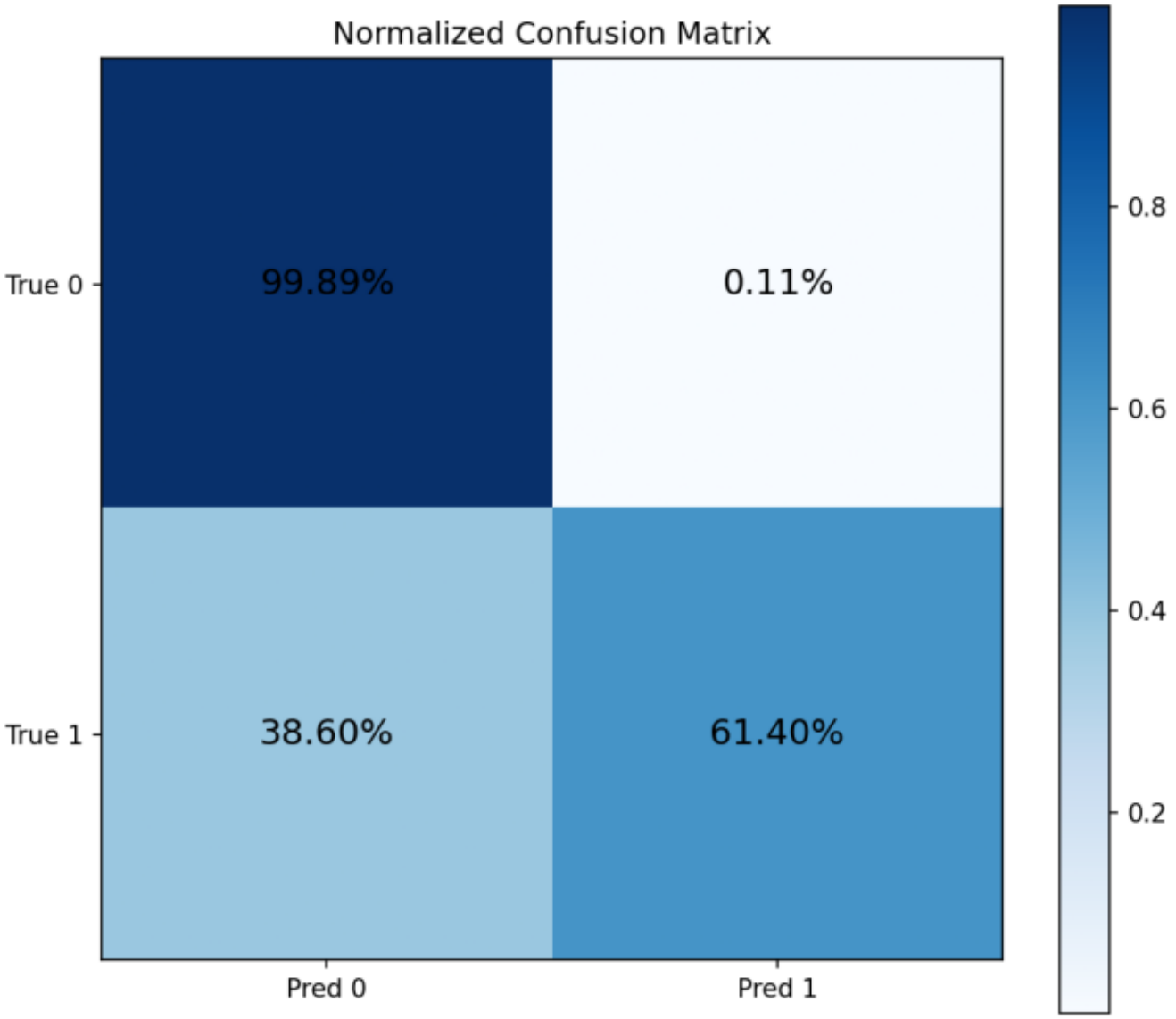
roc_curve: AUC-ROC=0.8426; curve above diagonal indicates strong separation.



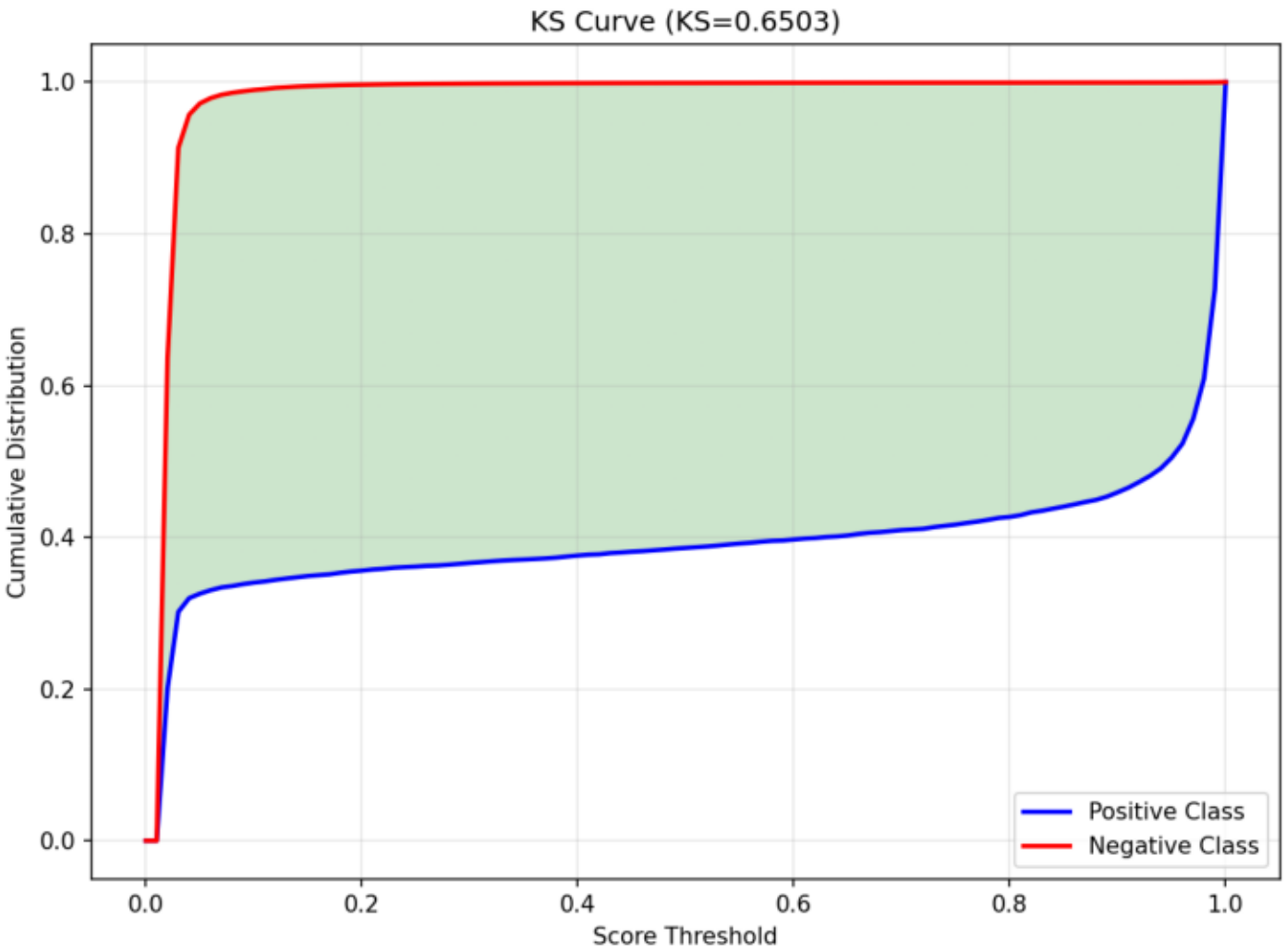
pr_curve: AUC-PR=0.7010 vs baseline 5.76%; higher curve implies better precision at recall.



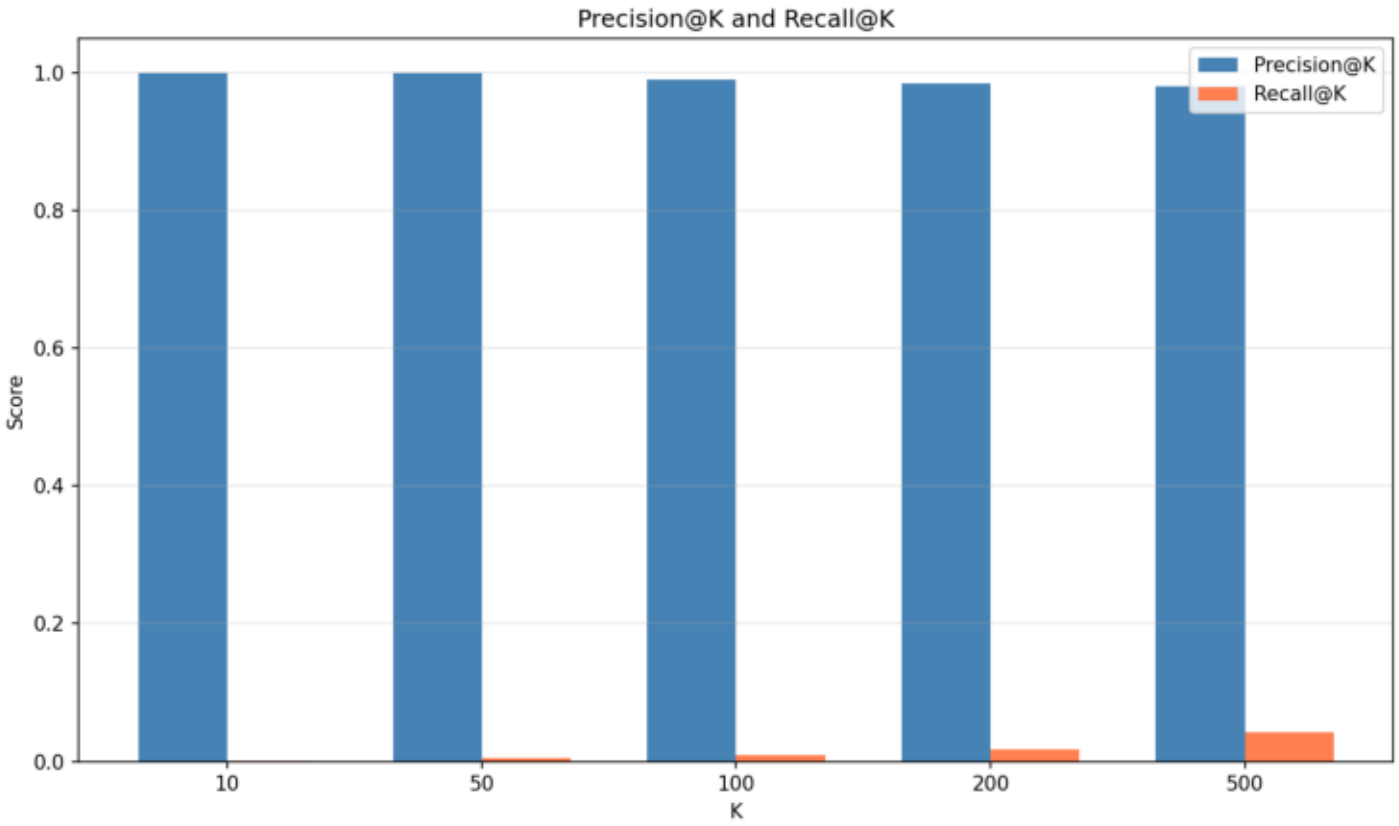
confusion_matrix: At threshold 0.50: TP=7072, FP=209, FN=4446, TN=188273.



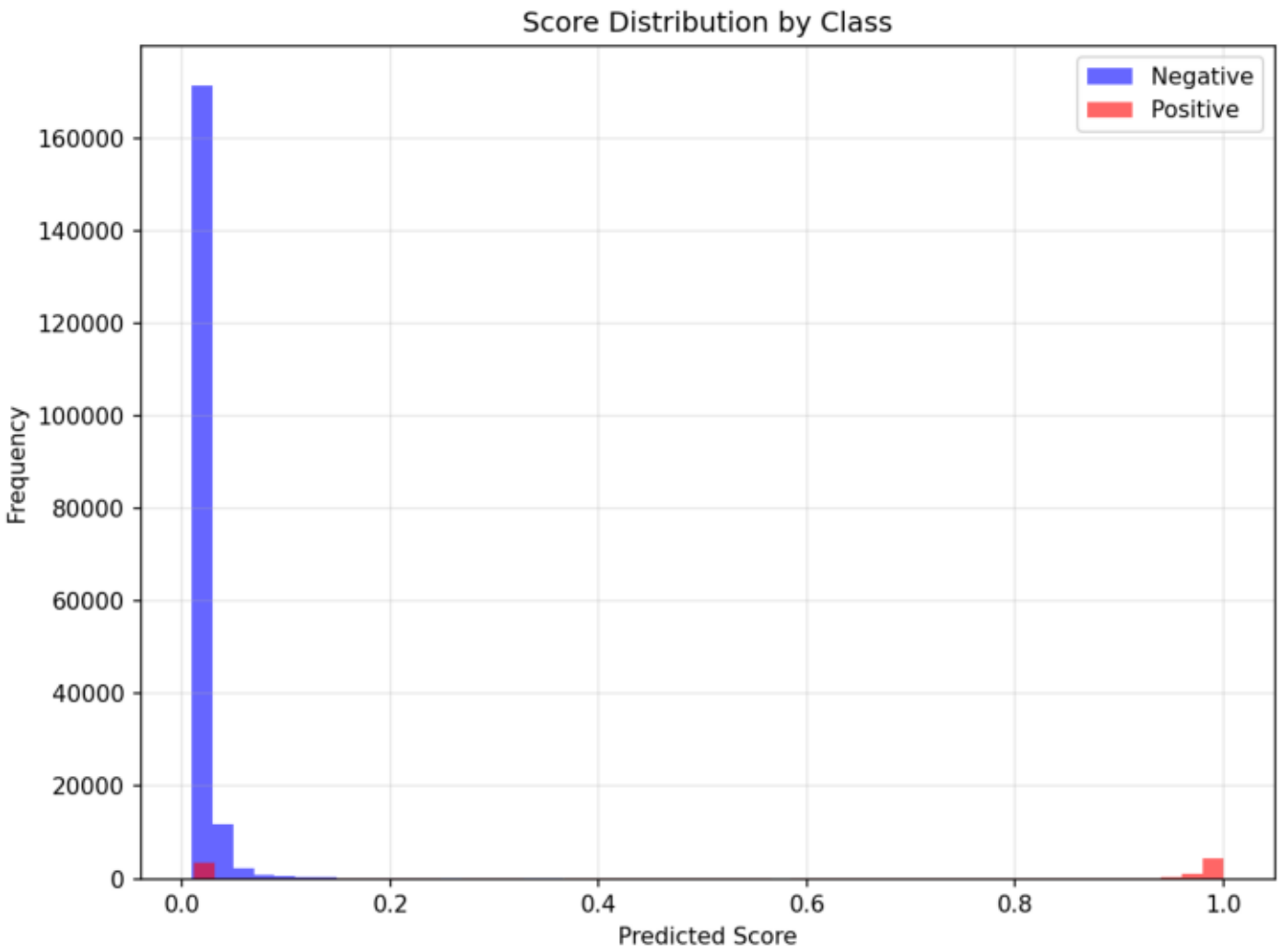
confusion_matrix_norm: Normalized rates: TPR=61.40%, TNR=99.89%.



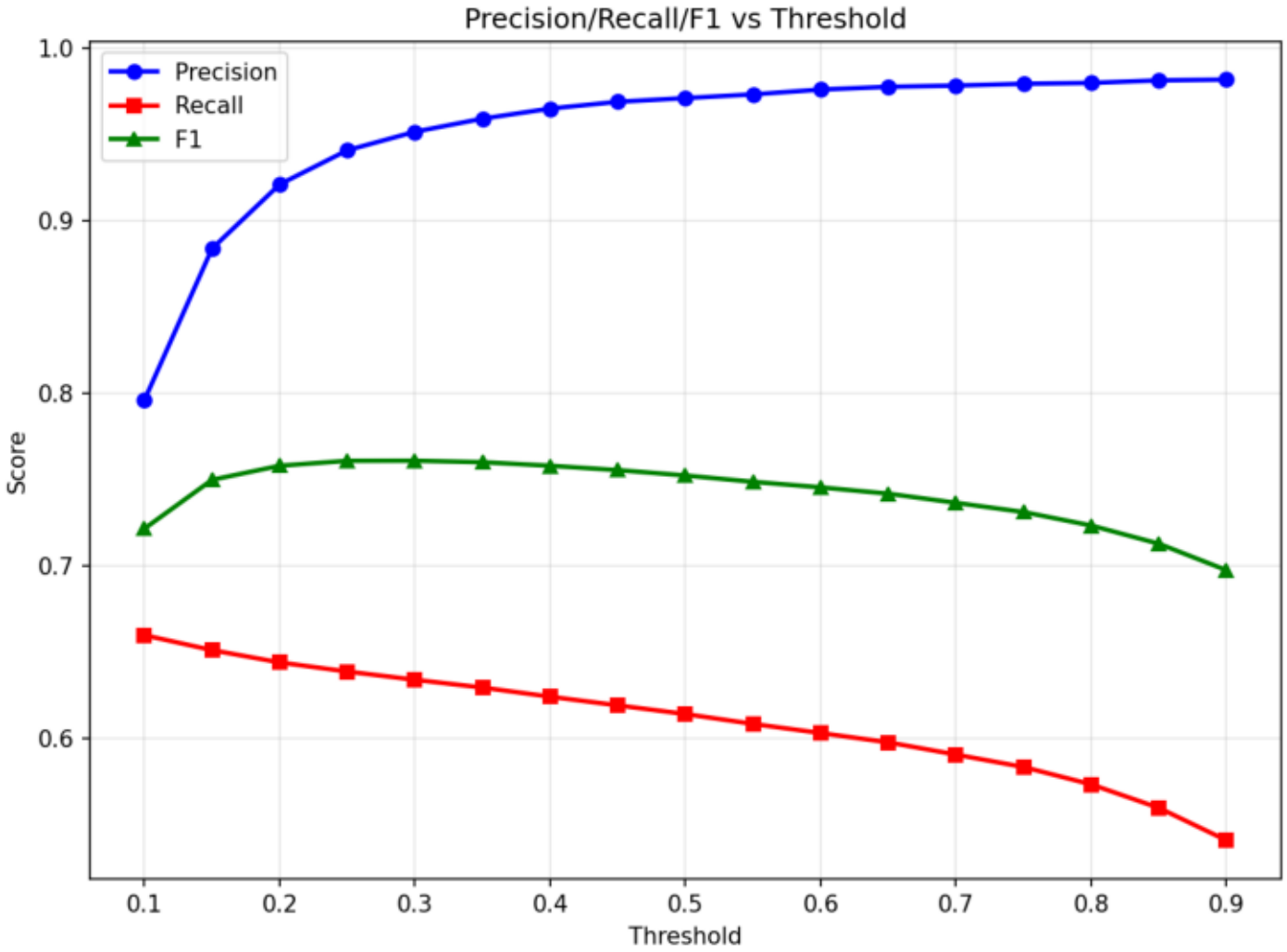
ks_curve: Maximum separation KS=0.6503 at threshold 0.08.



precision_recall_at_k: Bars show precision/recall as you expand the review queue.



score_distribution: Mean score: positive=0.607, negative=0.024 (larger gap is better).



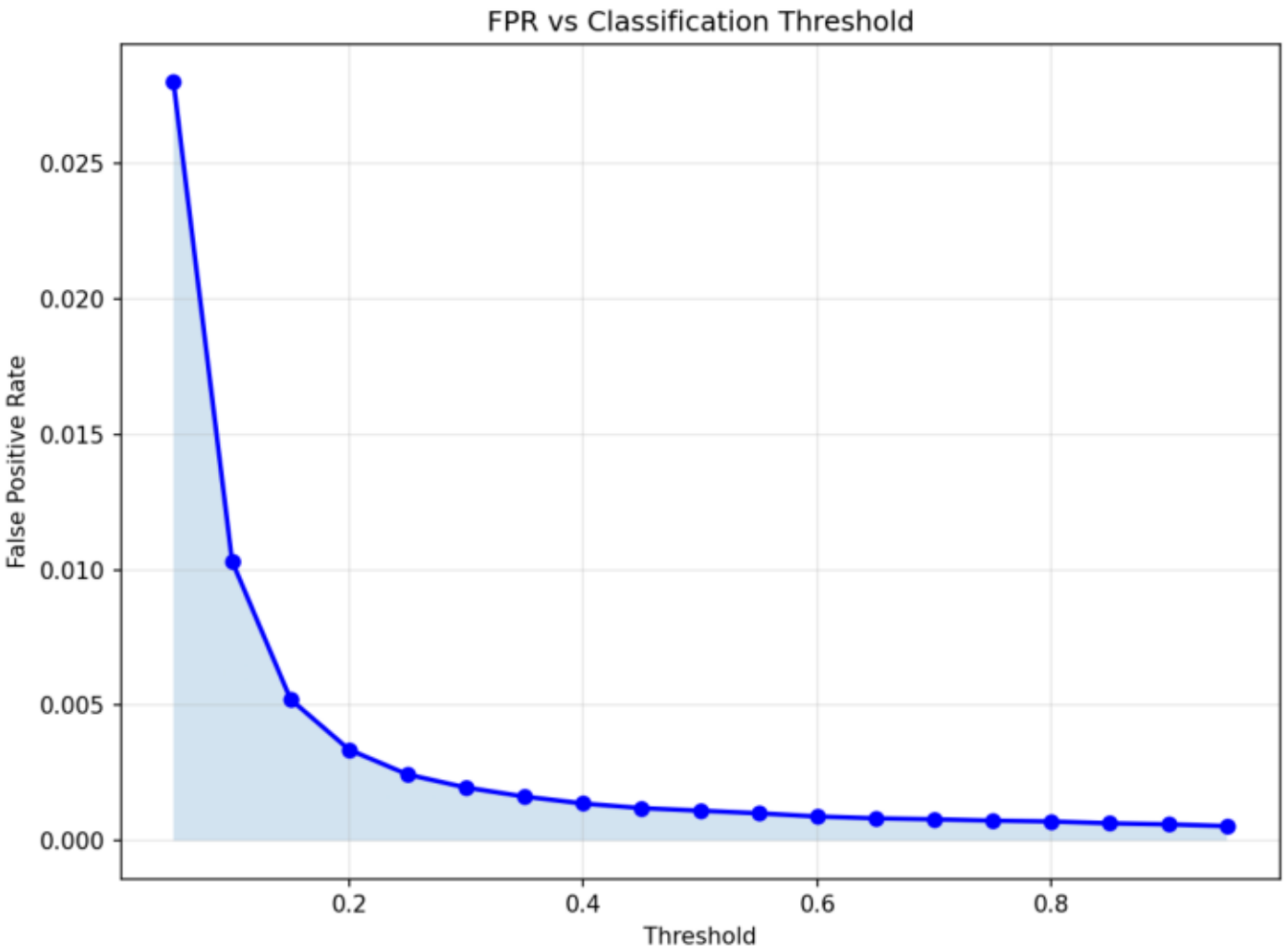
threshold_analysis: Best F1 in tested grid is 0.7610 at threshold 0.30.

2) Model Efficiency

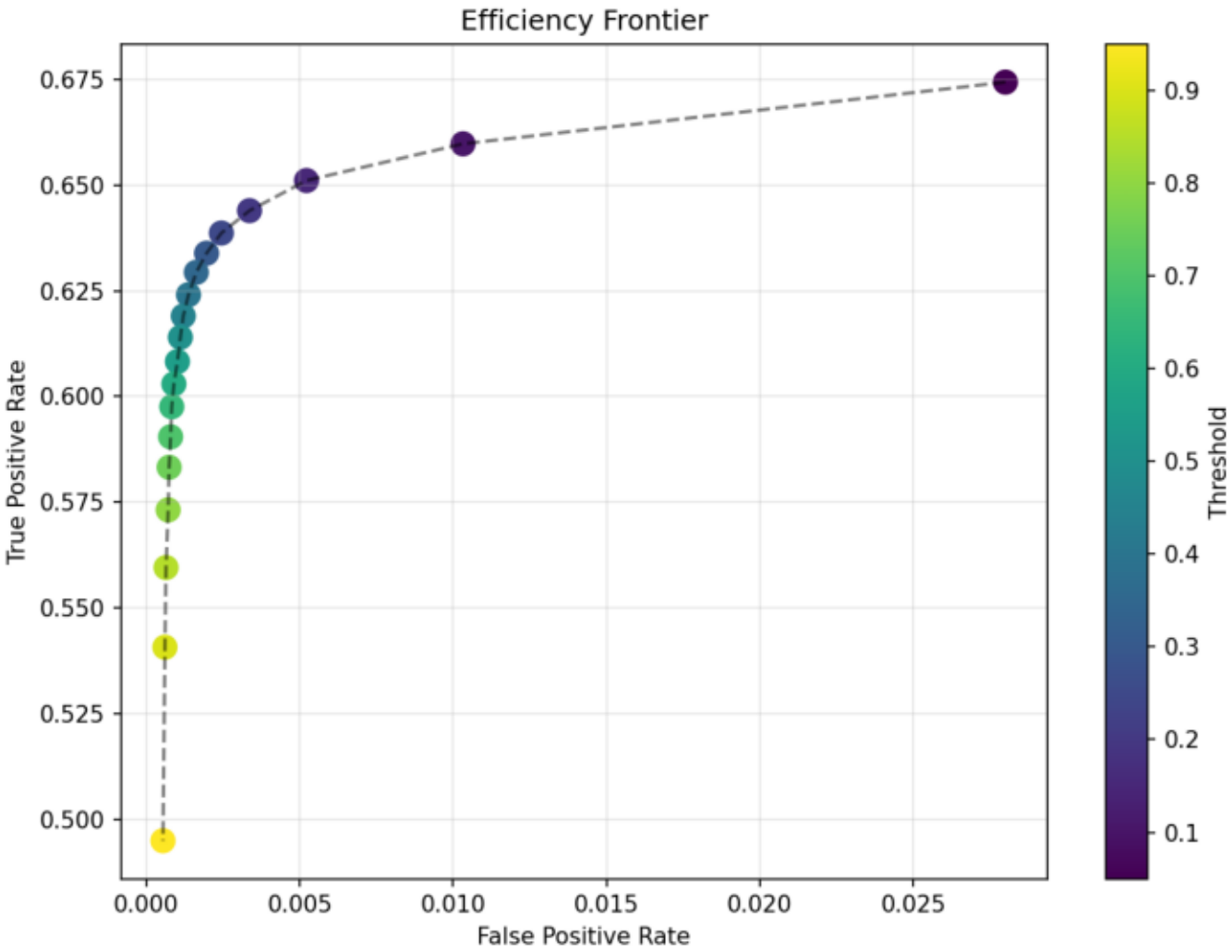
- fpr: 0.0011
- tn: 188273
- fp: 209
- threshold: 0.5000
- fpr_at_thresholds.t_0.05: 0.0280
- fpr_at_thresholds.t_0.10: 0.0103
- fpr_at_thresholds.t_0.15: 0.0052
- fpr_at_thresholds.t_0.20: 0.0034
- fpr_at_thresholds.t_0.25: 0.0024
- fpr_at_thresholds.t_0.30: 0.0020
- fpr_at_thresholds.t_0.35: 0.0016
- fpr_at_thresholds.t_0.40: 0.0014
- fpr_at_thresholds.t_0.45: 0.0012
- fpr_at_thresholds.t_0.50: 0.0011
- fpr_at_thresholds.t_0.55: 0.0010
- fpr_at_thresholds.t_0.60: 0.0009
- fpr_at_thresholds.t_0.65: 0.0008
- fpr_at_thresholds.t_0.70: 0.0008
- fpr_at_thresholds.t_0.75: 0.0007
- fpr_at_thresholds.t_0.80: 0.0007
- fpr_at_thresholds.t_0.85: 0.0006
- fpr_at_thresholds.t_0.90: 0.0006
- fpr_at_thresholds.t_0.95: 0.0005

2) Model Efficiency - Explanations

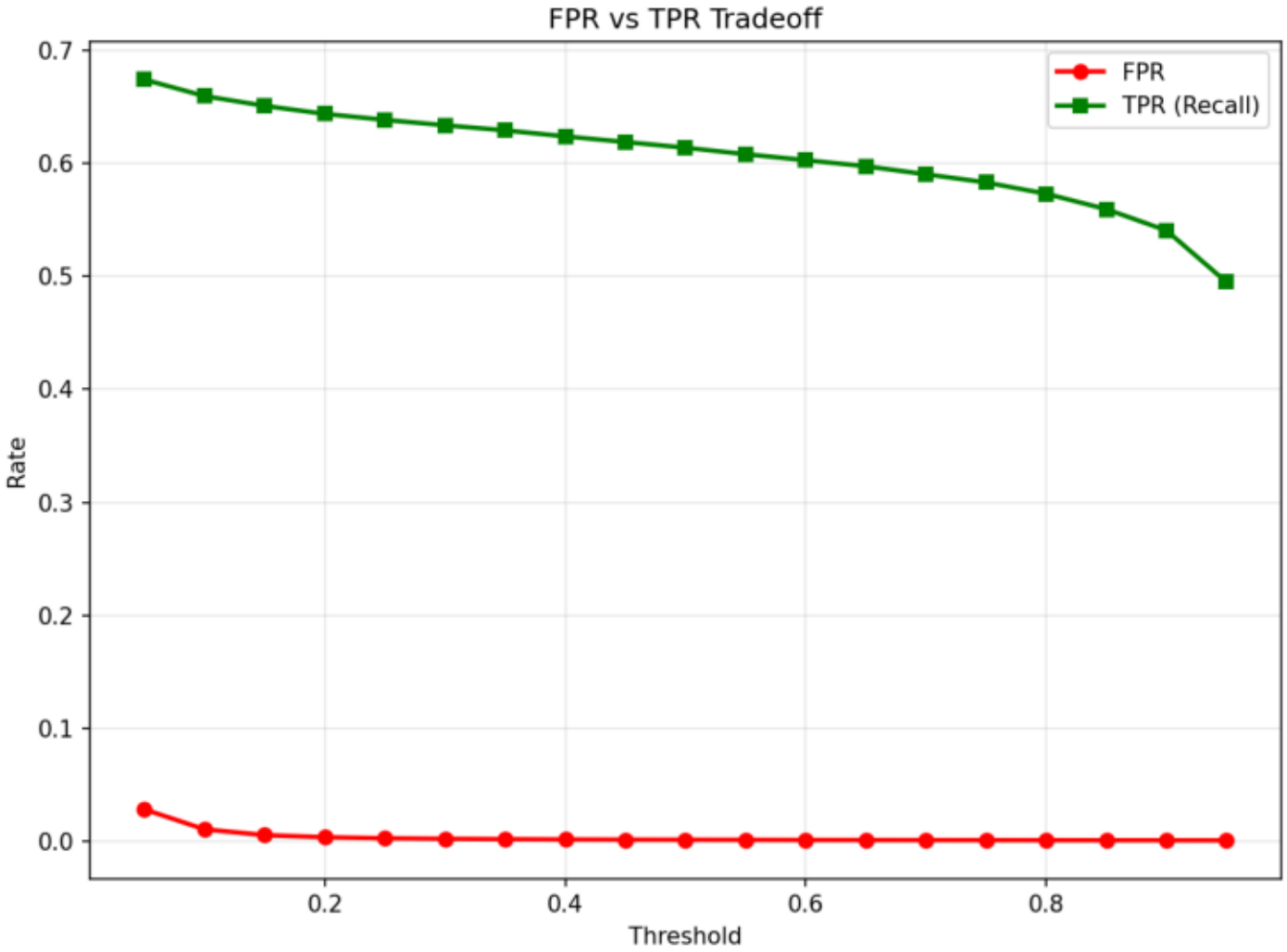
- At threshold 0.50, FPR=0.0011 (low); FP=209 out of 188482 negatives.
- At the same threshold, TPR=0.6140 with TP=7072 and FN=4446, showing the capture rate of positives.
- A threshold near 0.05 yields FPR≈0.0280 with TPR≈0.6743 if you want to target ~5% false positives.



fpr_vs_threshold: FPR decreases as the threshold increases; use it to pick an operating point.



efficiency_frontier: Each point shows the FPR/TPR tradeoff; move toward the top-left for better efficiency.



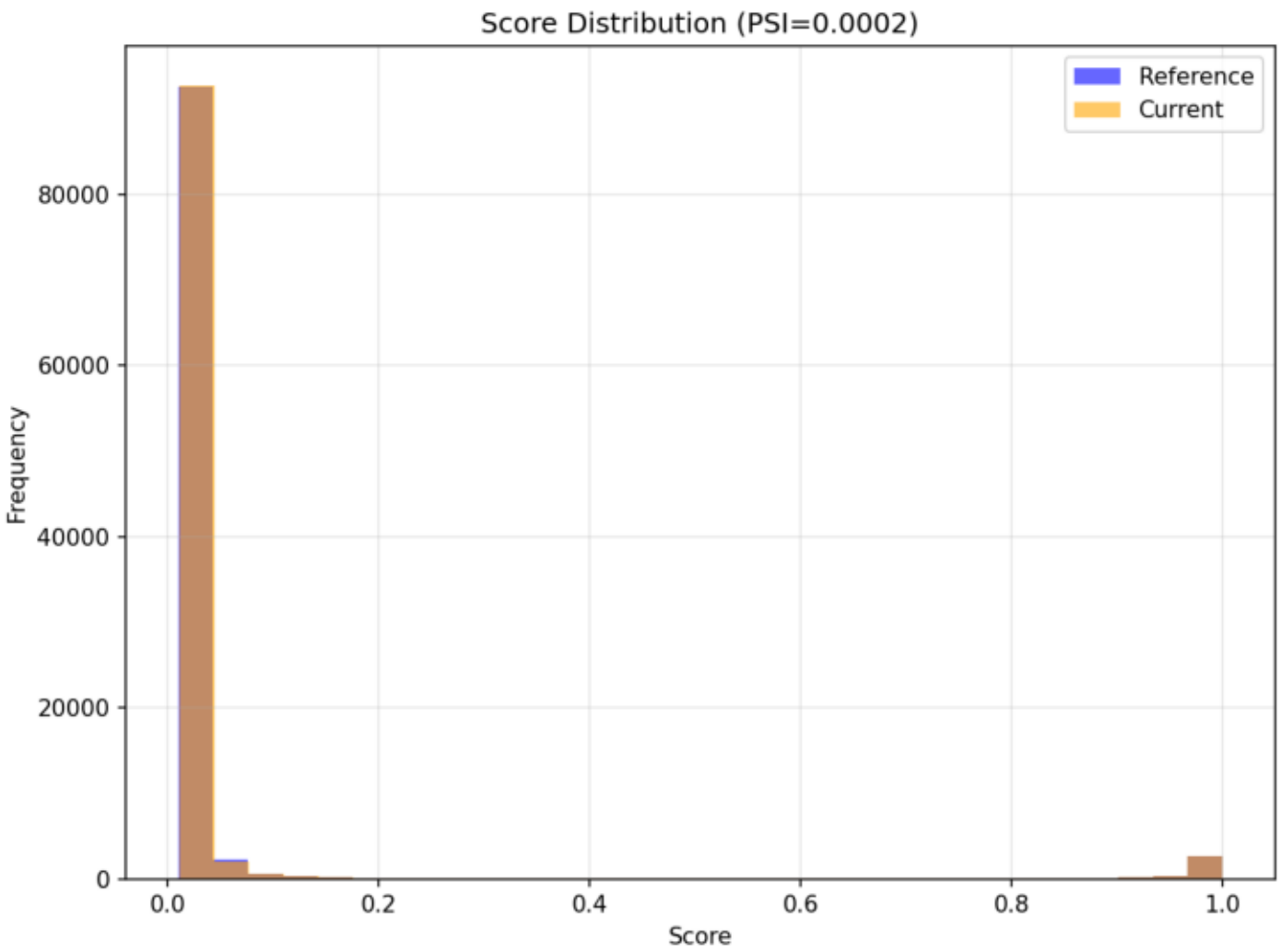
fpr_tpr_tradeoff: FPR and TPR curves highlight how recall drops as you reduce false positives.

3) Model Stability

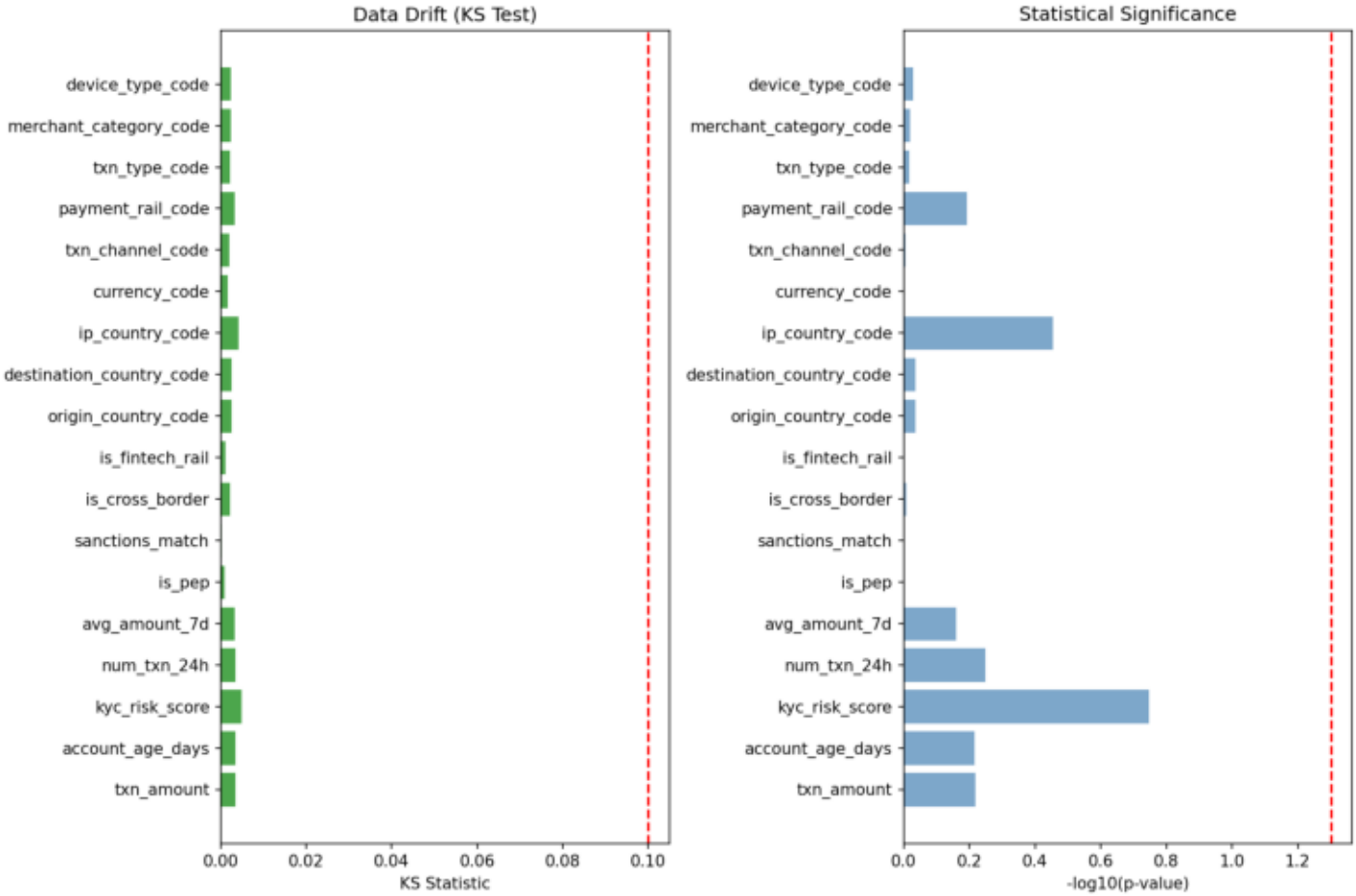
- psi: 0.0002
- cv_auc_roc_mean: 0.8426
- cv_auc_roc_std: 0.0047
- cv_auc_pr_mean: 0.7010
- cv_auc_pr_std: 0.0086
- bootstrap_auc_roc_mean: 0.8400
- bootstrap_auc_roc_ci_lower: 0.8338
- bootstrap_auc_roc_ci_upper: 0.8474
- concept_drift_detected: False
- concept_drift_score: 0.0078

3) Model Stability - Explanations

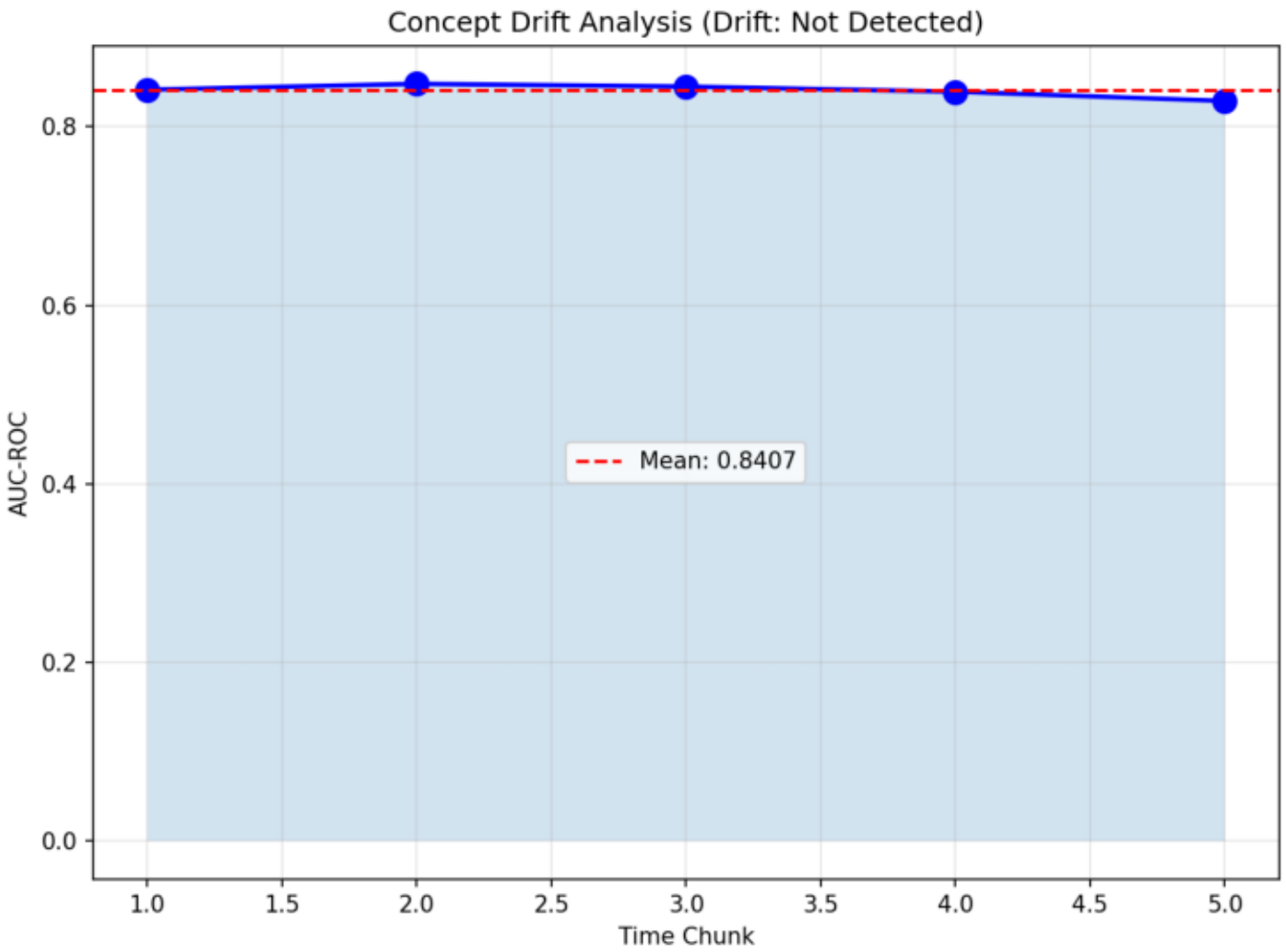
- PSI is 0.0002, indicating negligible score distribution shift between reference and current data.
- CV AUC-ROC is 0.8426 ± 0.0047 (stable across folds).
- CV AUC-PR is 0.7010 ± 0.0086 (stable across folds).
- Bootstrap AUC-ROC mean is 0.8400 with 95% CI [0.8338, 0.8474] (width 0.0136).
- No concept drift detected (score 0.0078), suggesting stable performance over time.
- Data drift flagged 0/18 numeric features; top drift signals: ['kyc_risk_score', 'ip_country_code',
- 'num_txn_24h'].



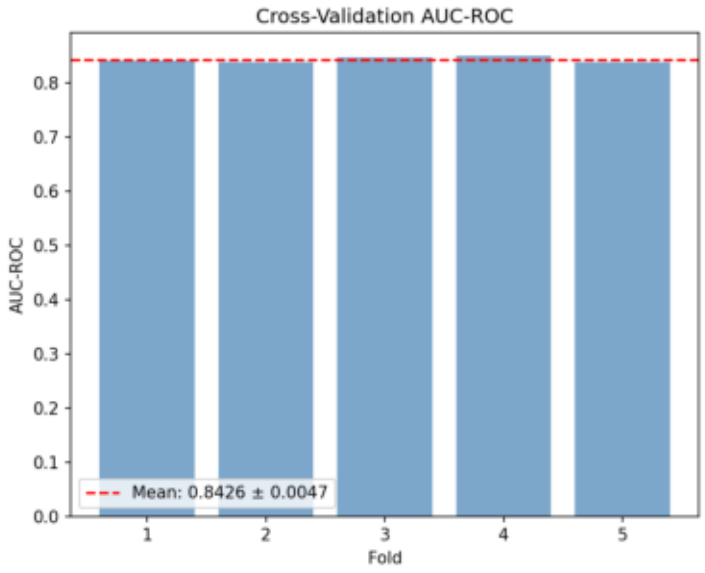
psi_distribution: Score distributions overlap consistent with PSI=0.0002.

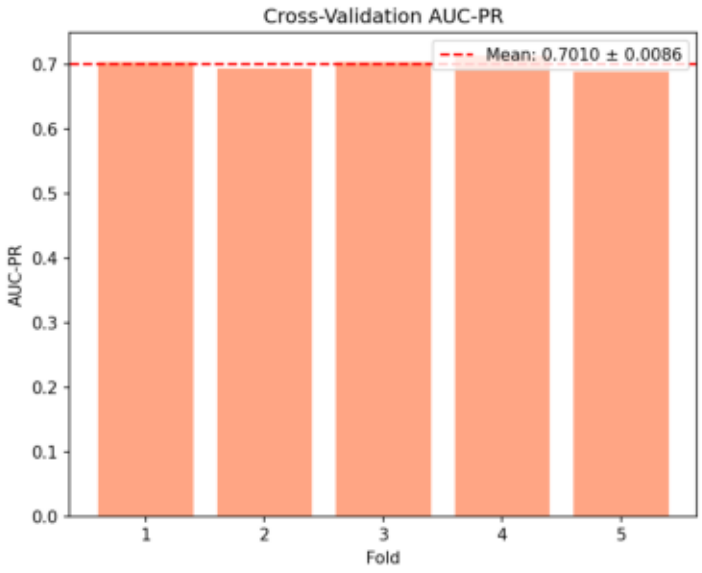


data_drift_heatmap: 0/18 features show drift; red bars highlight flagged features.

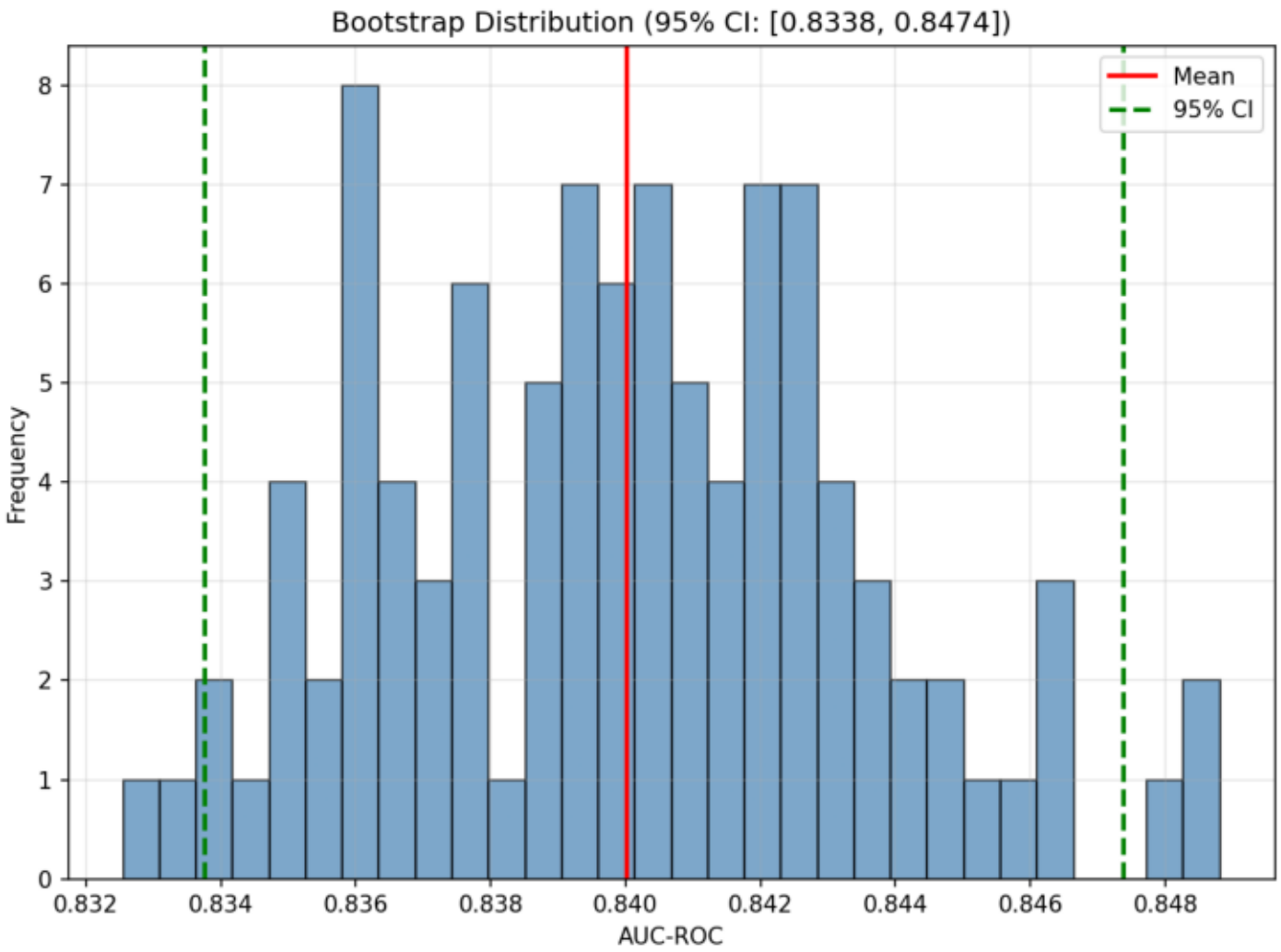


concept_drift: AUC across time chunks shows whether performance is stable over time.

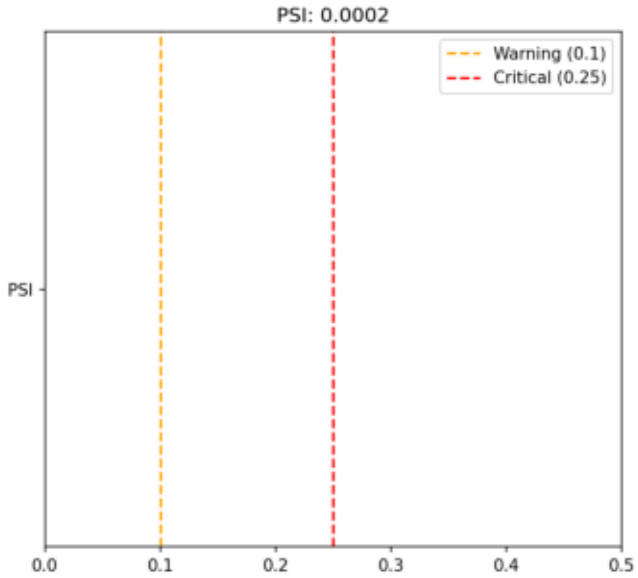


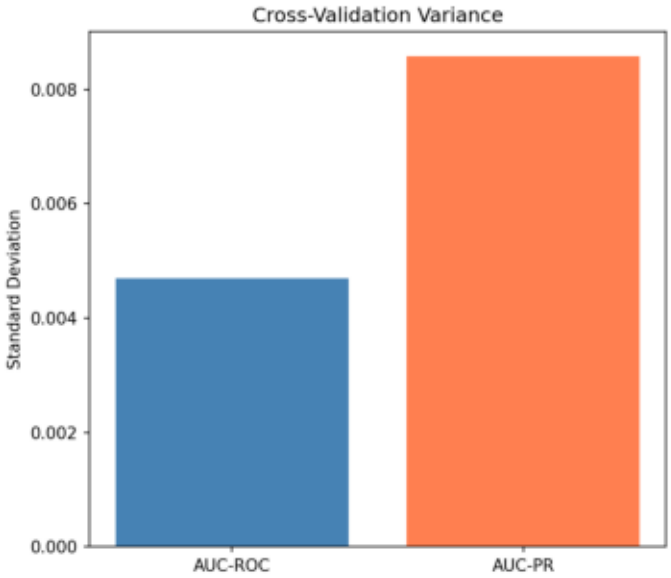


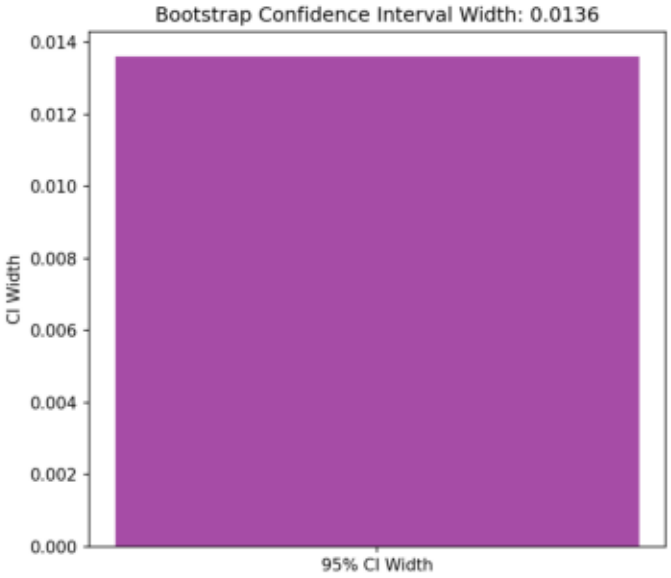
cv_results: Fold scores cluster around ROC 0.8426 and PR 0.7010.



bootstrap_distribution: Bootstrap CI [0.8338, 0.8474] reflects performance stability.








```
STABILITY SUMMARY
PSI: 0.0002 (OK)

CV AUC-ROC: 0.8426 ± 0.0047
CV AUC-PR:  0.7010 ± 0.0086

Bootstrap 95% CI: [0.8338, 0.8474]
```

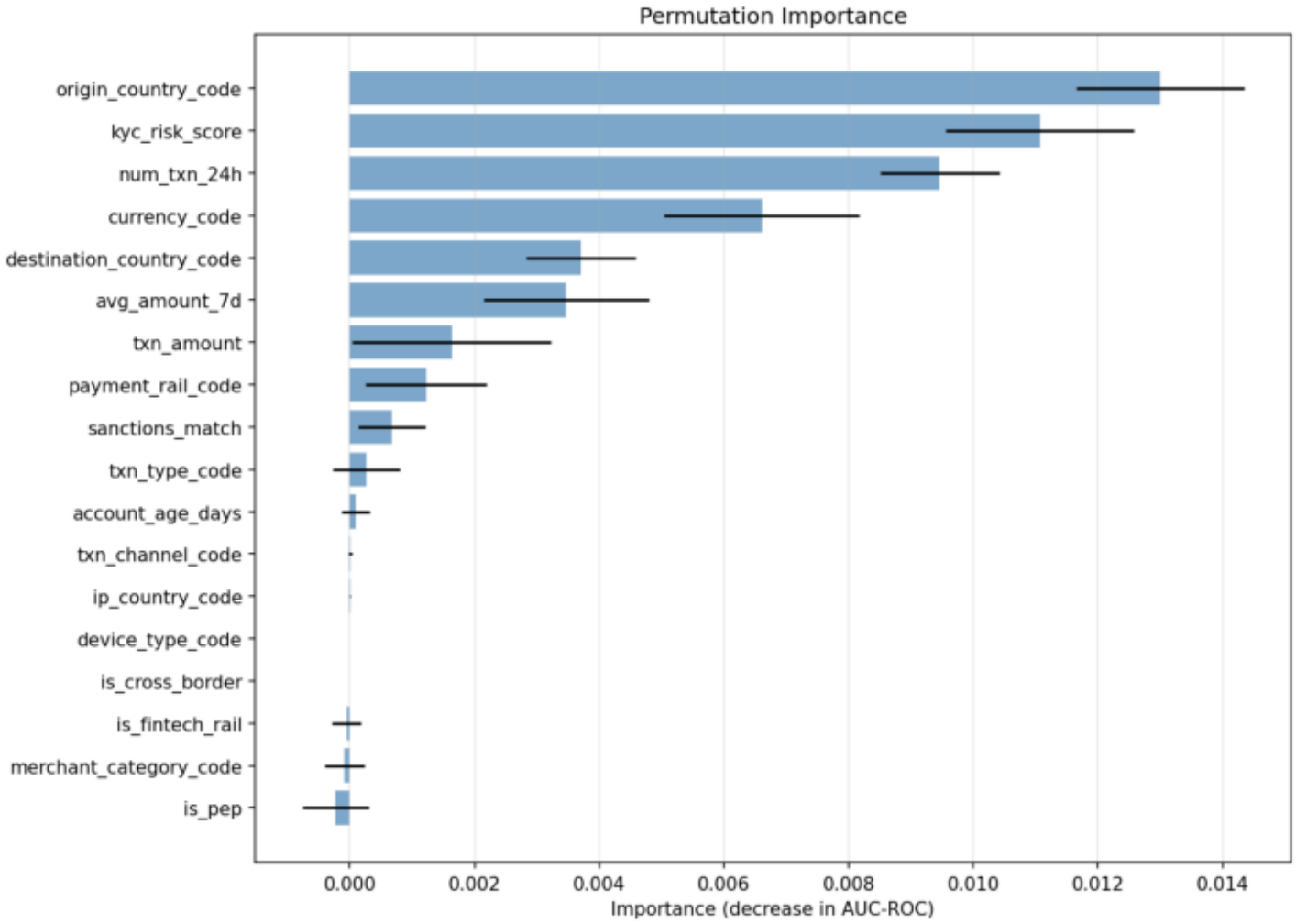
stability_summary: Dashboard summarizes PSI, CV variance, and bootstrap CI width.

4) Model Interpretability

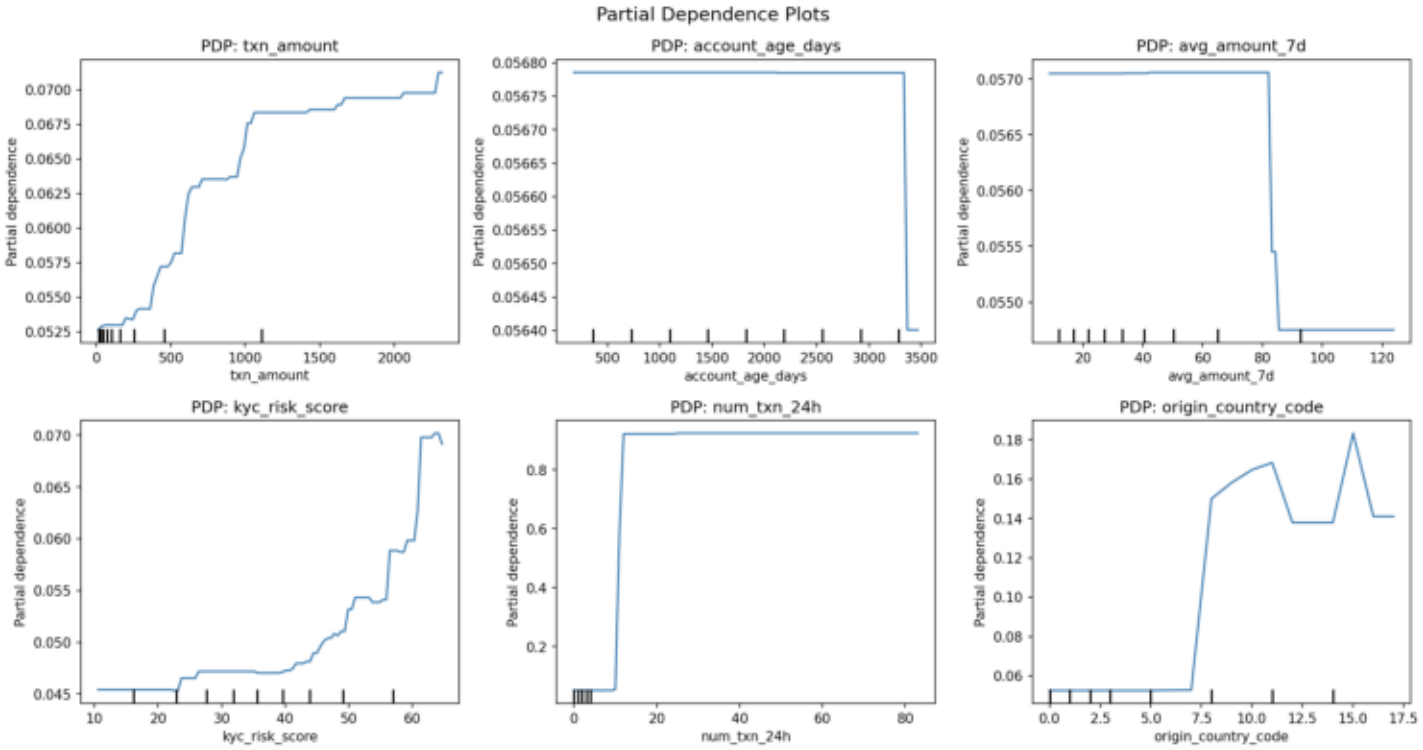
- model_type: tree
- methods_used: ['permutation', 'pdp', 'ice']
- shap_status: unavailable
- shap_reason: No module named 'shap'
- perm_top_features: ['origin_country_code', 'kyc_risk_score', 'num_txn_24h', 'currency_code', 'destination_country_code']
- pdp_features: ['txn_amount', 'account_age_days', 'avg_amount_7d', 'kyc_risk_score', 'num_txn_24h']...
- ice_features: ['txn_amount', 'account_age_days', 'avg_amount_7d', 'kyc_risk_score']

4) Model Interpretability - Explanations

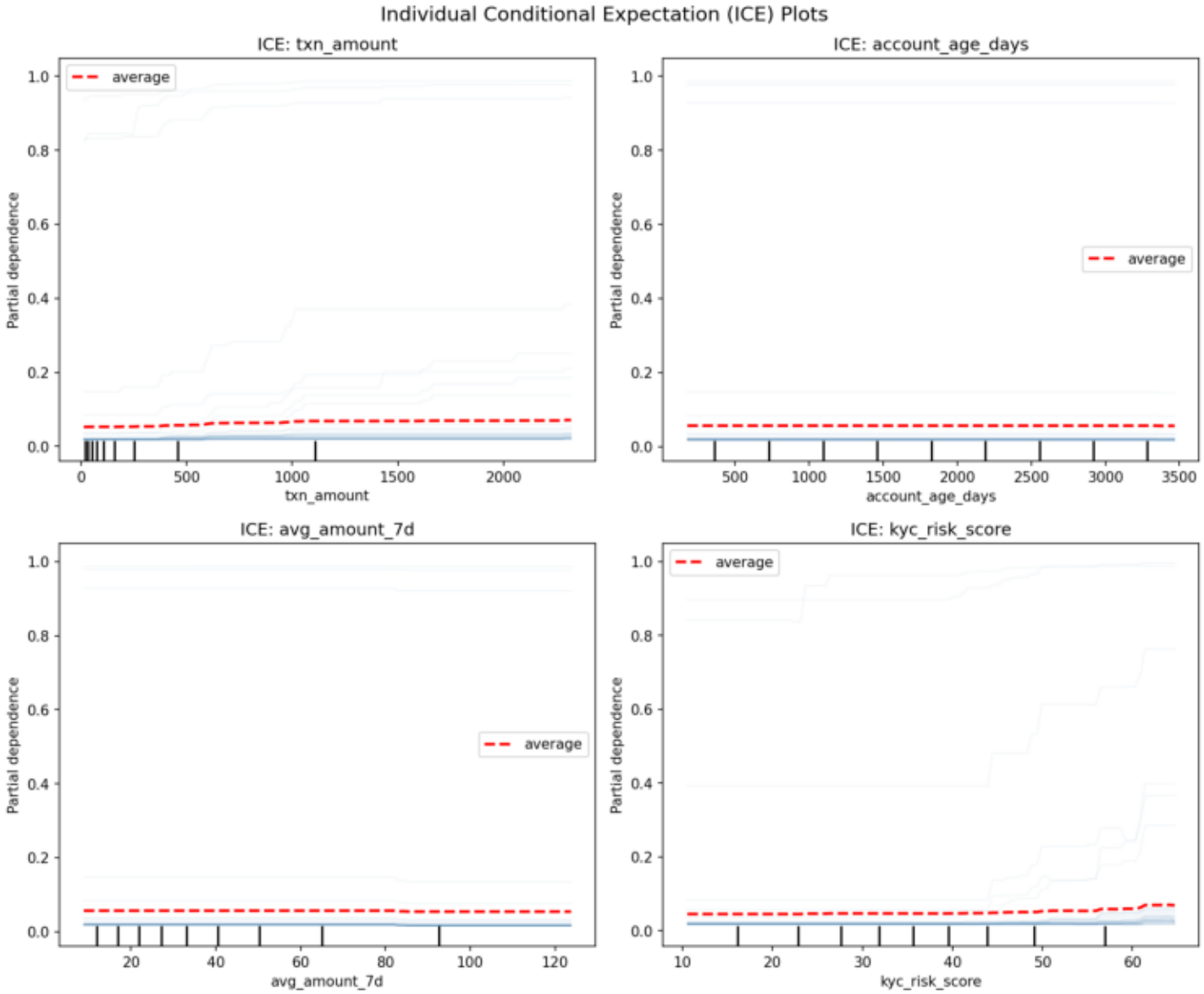
- Permutation importance highlights: ['origin_country_code', 'kyc_risk_score', 'num_txn_24h', 'currency_code',
- 'destination_country_code'].
- SHAP plots not generated because SHAP is unavailable (No module named 'shap').
- PDP plots show average effects for: ['txn_amount', 'account_age_days', 'avg_amount_7d', 'kyc_risk_score',
- 'num_txn_24h', 'origin_country_code'].
- ICE plots show individual-level effects for: ['txn_amount', 'account_age_days', 'avg_amount_7d',
- 'kyc_risk_score'].



permutation_importance: Bars show how much AUC drops when each feature is permuted.



pdp: PDP curves show average model response as a feature varies.



ice: ICE curves show per-instance responses as a feature varies.