Model Testing Report

Experiment: model_testing

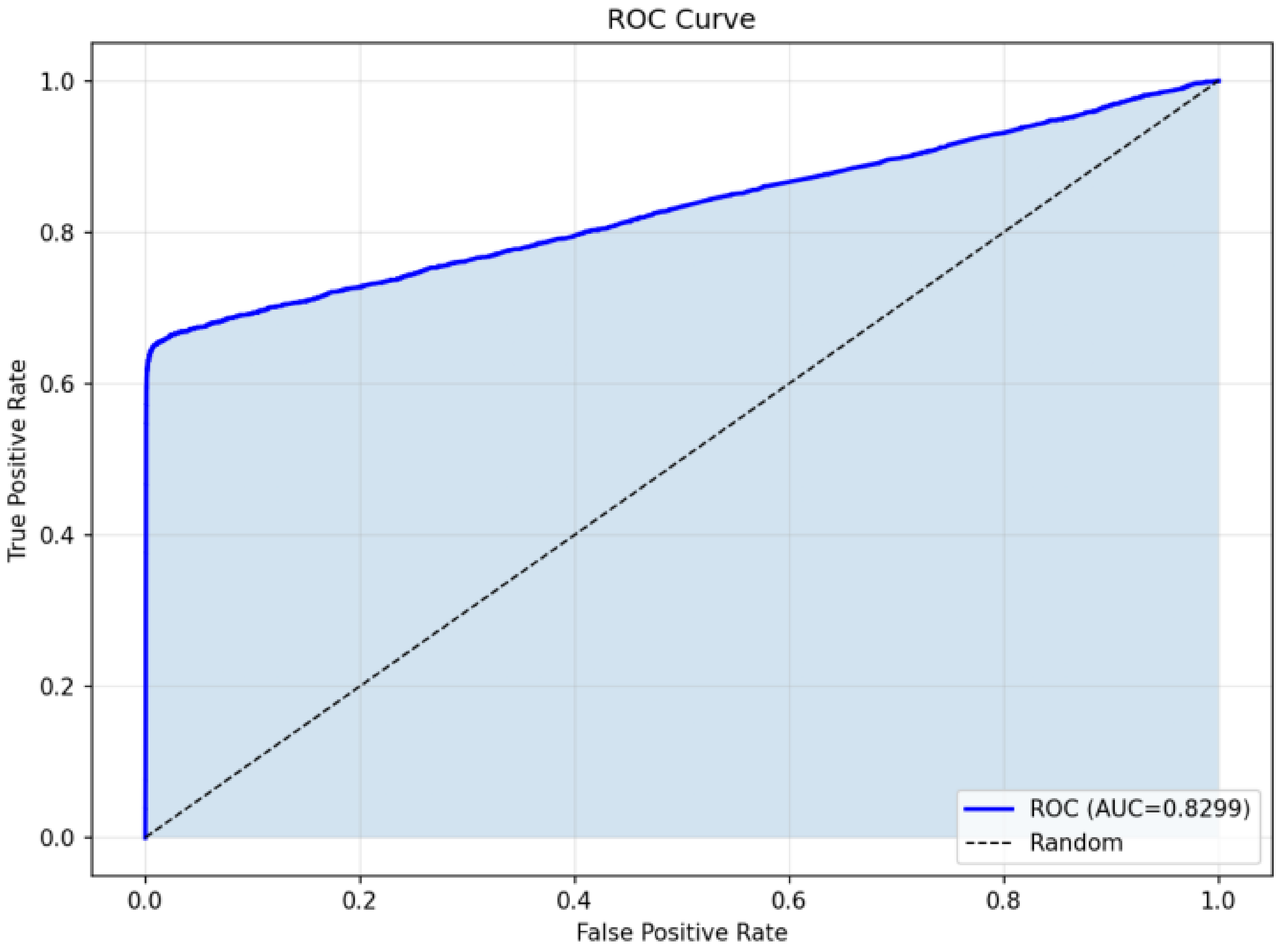
Generated: 2026-01-28 15:39:16

1) Model Effectiveness

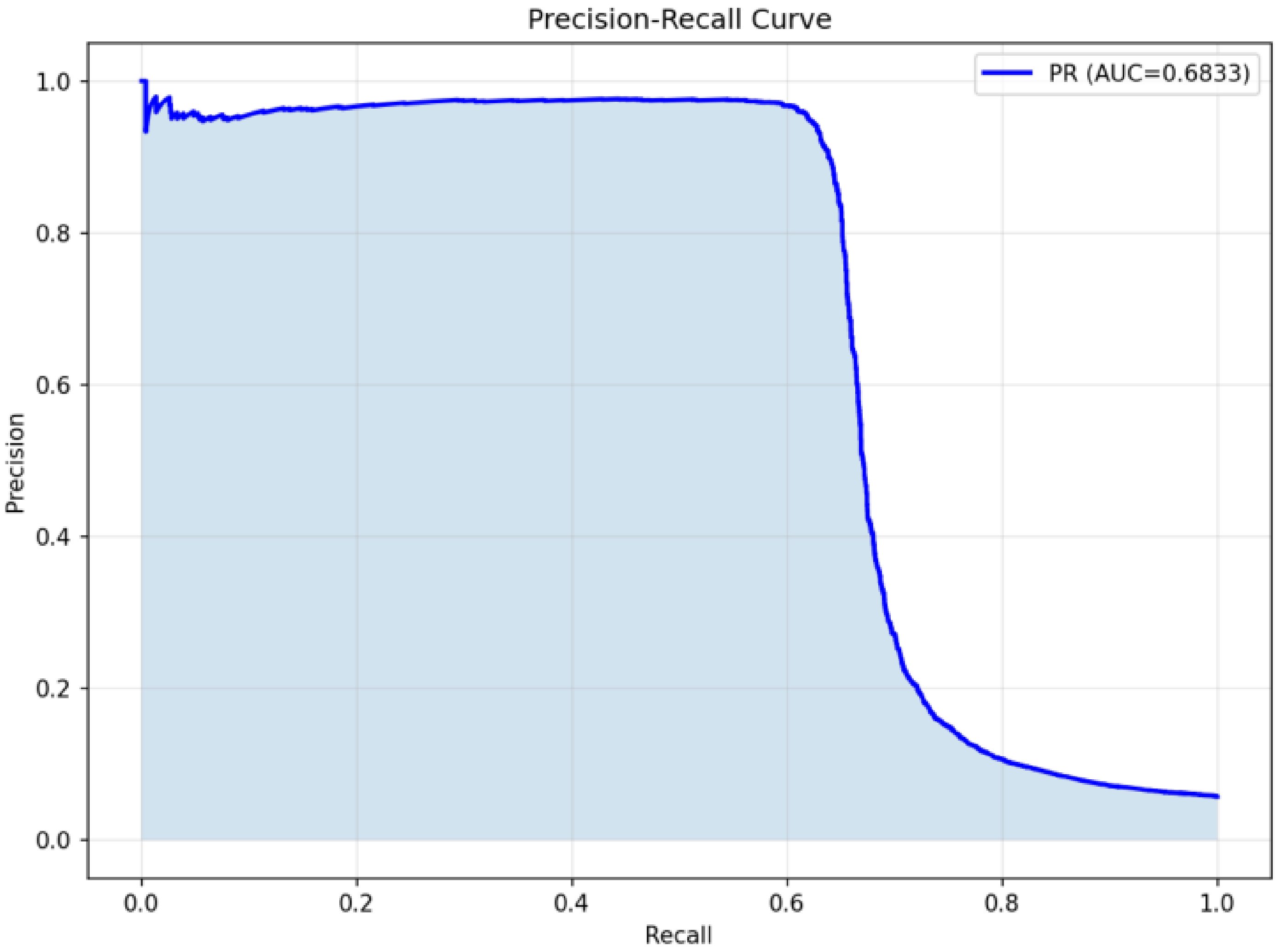
- auc_roc: 0.8299
- auc_pr: 0.6833
- precision: 0.9666
- recall: 0.6035
- f1: 0.7431
- ks_statistic: 0.6422
- ks_threshold: 0.1200
- confusion_matrix.TN: 56473
- confusion_matrix.FP: 72
- confusion_matrix.FN: 1370
- confusion_matrix.TP: 2085
- precision_at_k.10: 1.0000
- precision_at_k.50: 0.9600
- precision_at_k.100: 0.9600
- precision_at_k.200: 0.9500
- precision_at_k.500: 0.9620
- recall_at_k.10: 0.0029
- recall_at_k.50: 0.0139
- recall_at_k.100: 0.0278
- recall_at_k.200: 0.0550
- recall_at_k.500: 0.1392

1) Model Effectiveness - Explanations

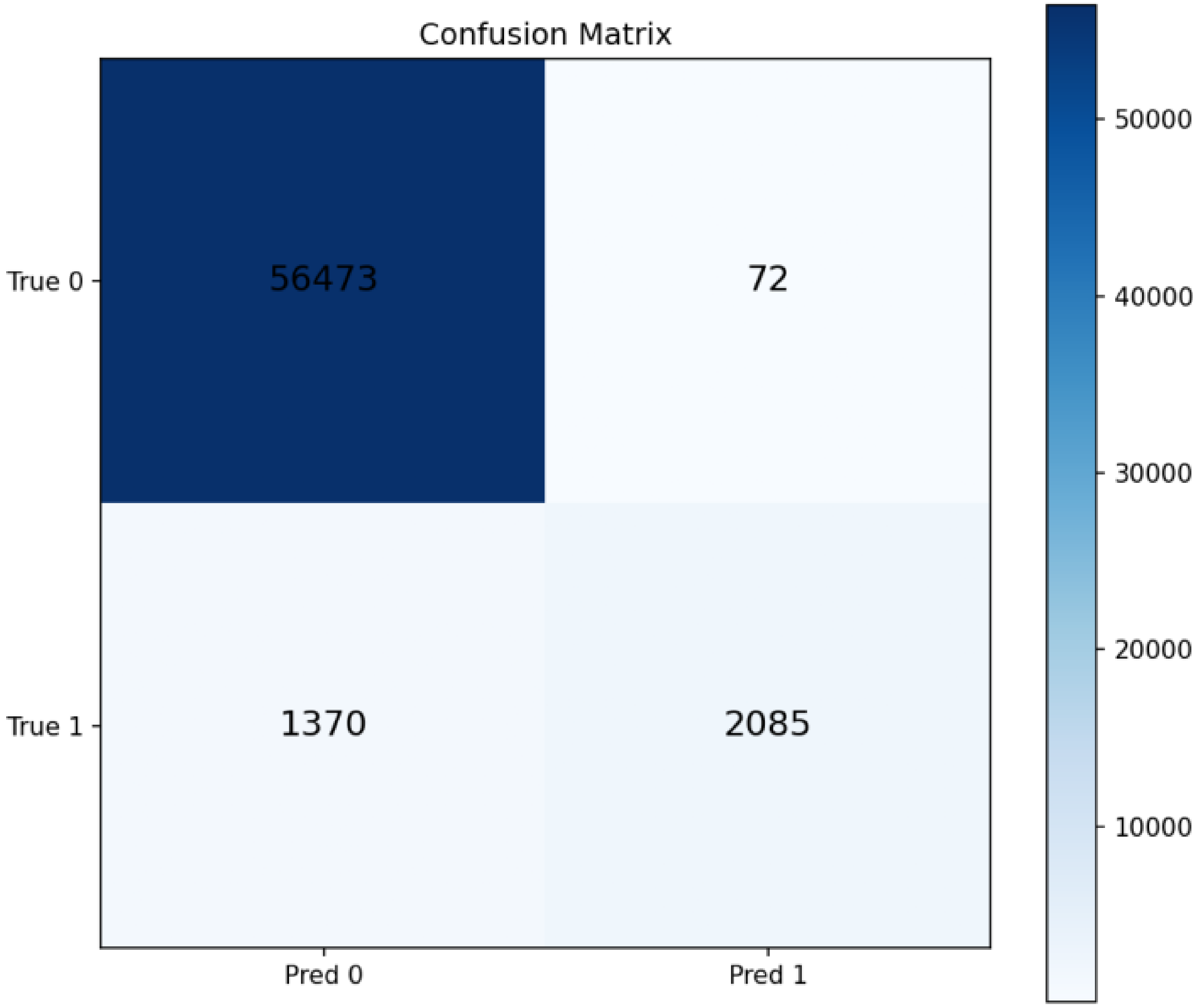
- Positive rate is 5.76%; AUC-ROC is 0.8299 (strong ranking).
- AUC-PR is 0.6833, above the base rate 5.76%, indicating meaningful lift.
- At threshold 0.50, precision=0.9666, recall=0.6035, F1=0.7431 (precision-heavy (fewer positives caught, fewer
- false alarms)).
- Confusion matrix: TP=2085, FP=72, FN=1370, TN=56473 (FPR=0.13%, FNR=39.65%).
- KS is 0.6422 at threshold 0.12, indicating very strong separation.
- Top-10: precision=1.0000, recall=0.0029 shows the quality of the highest-risk shortlist.
- Top-500: precision=0.9620, recall=0.1392 shows how much coverage you get with a larger queue.



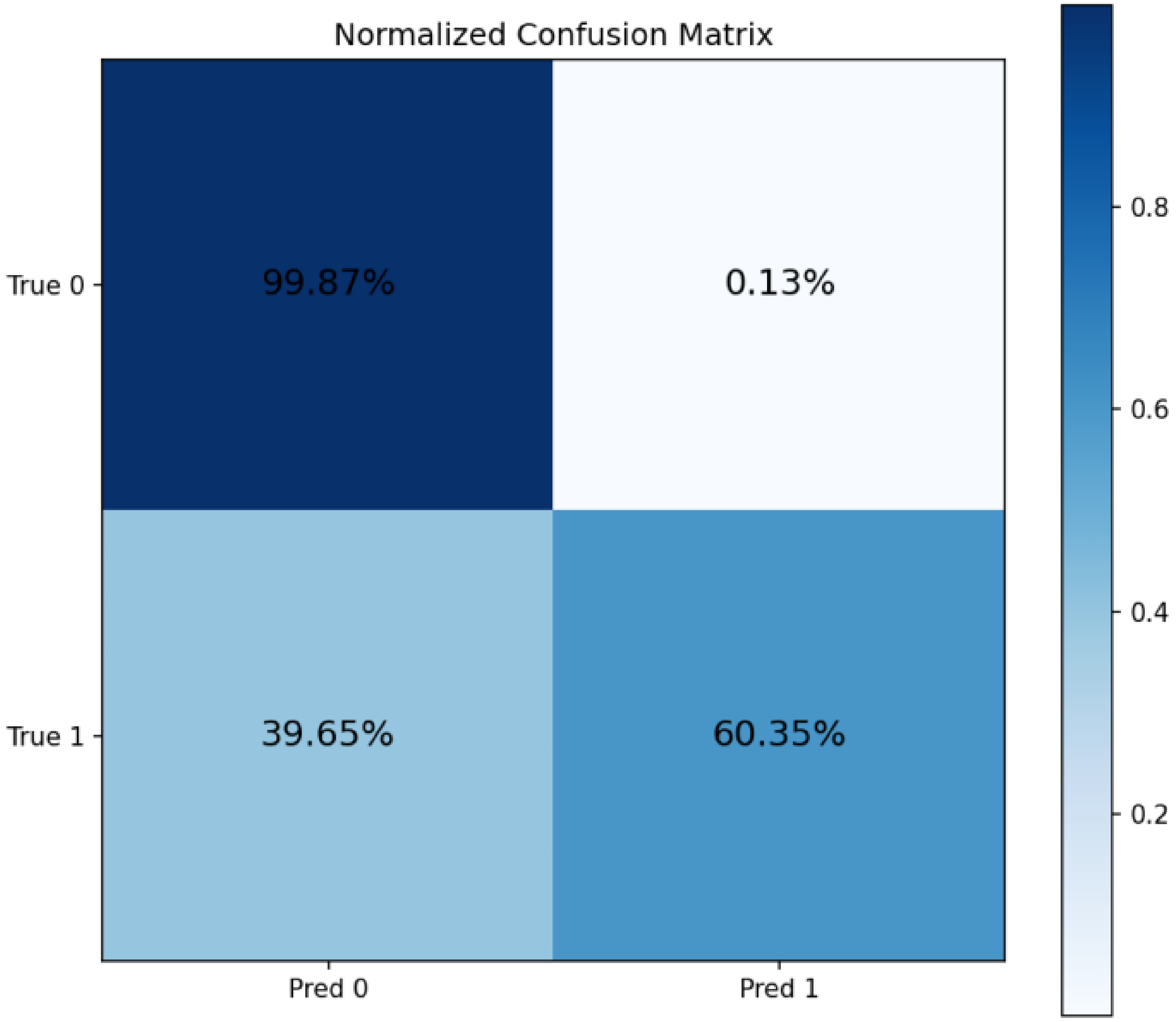
roc_curve: AUC-ROC=0.8299; curve above diagonal indicates strong separation.



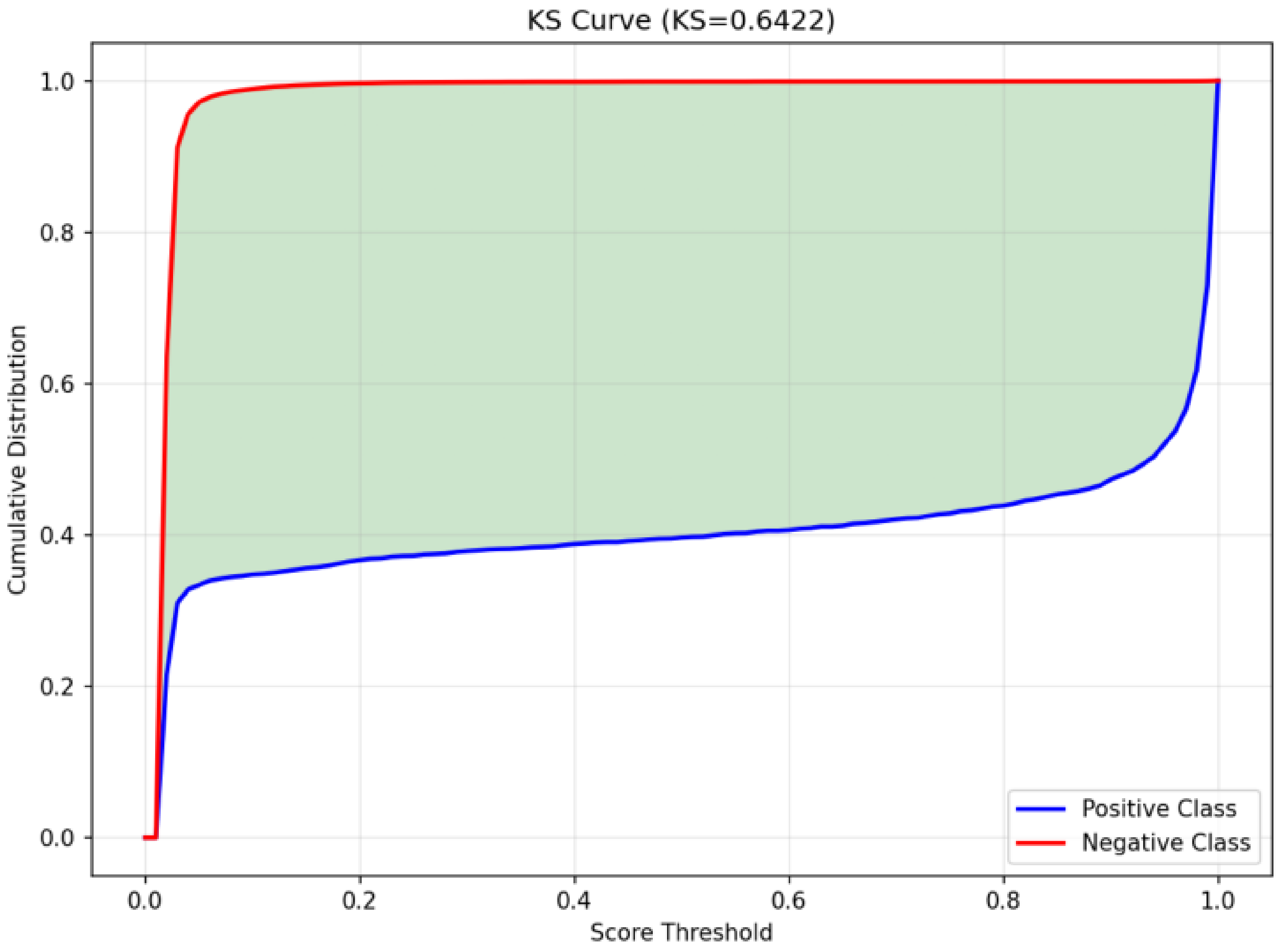
pr_curve: AUC-PR=0.6833 vs baseline 5.76%; higher curve implies better precision at recall.



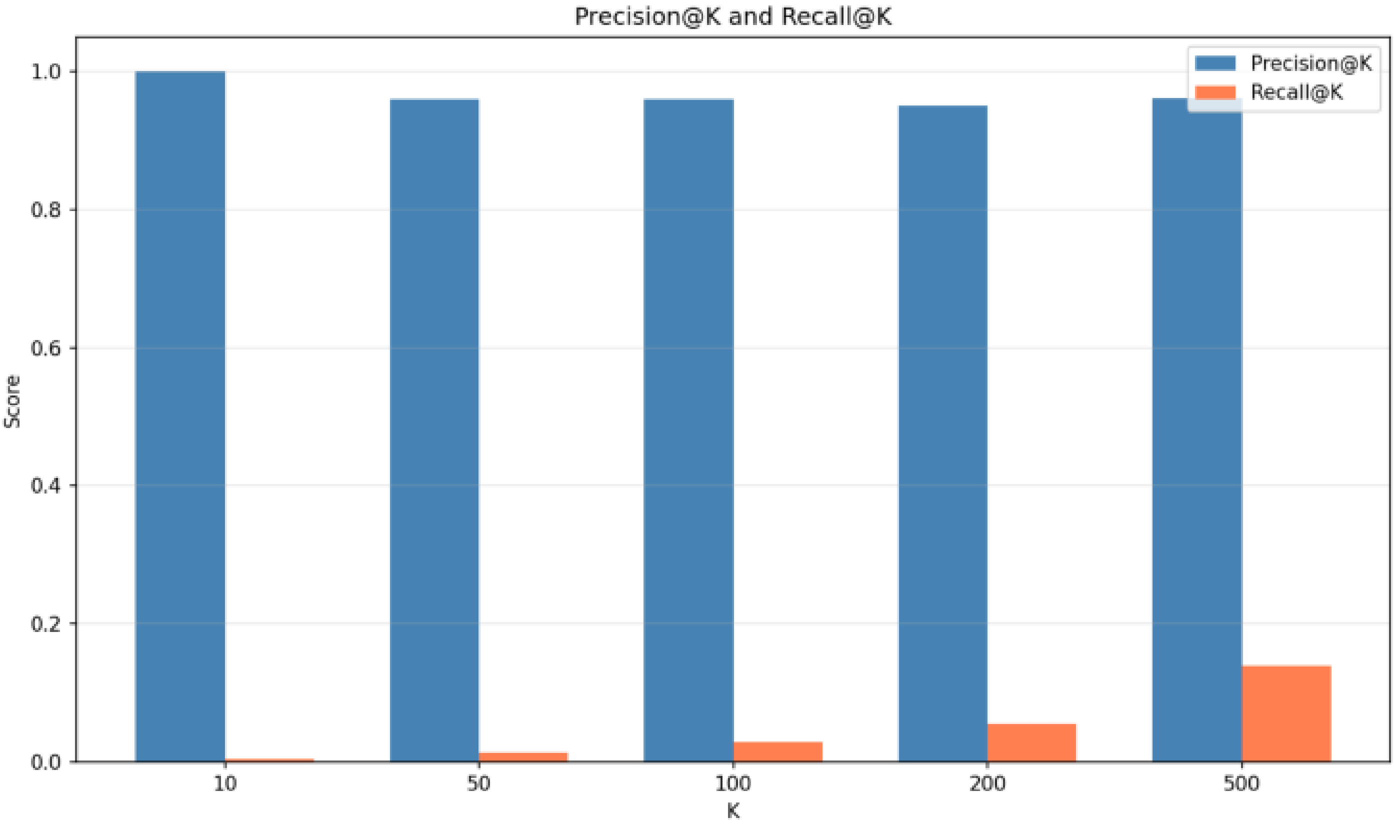
confusion_matrix: At threshold 0.50: TP=2085, FP=72, FN=1370, TN=56473.



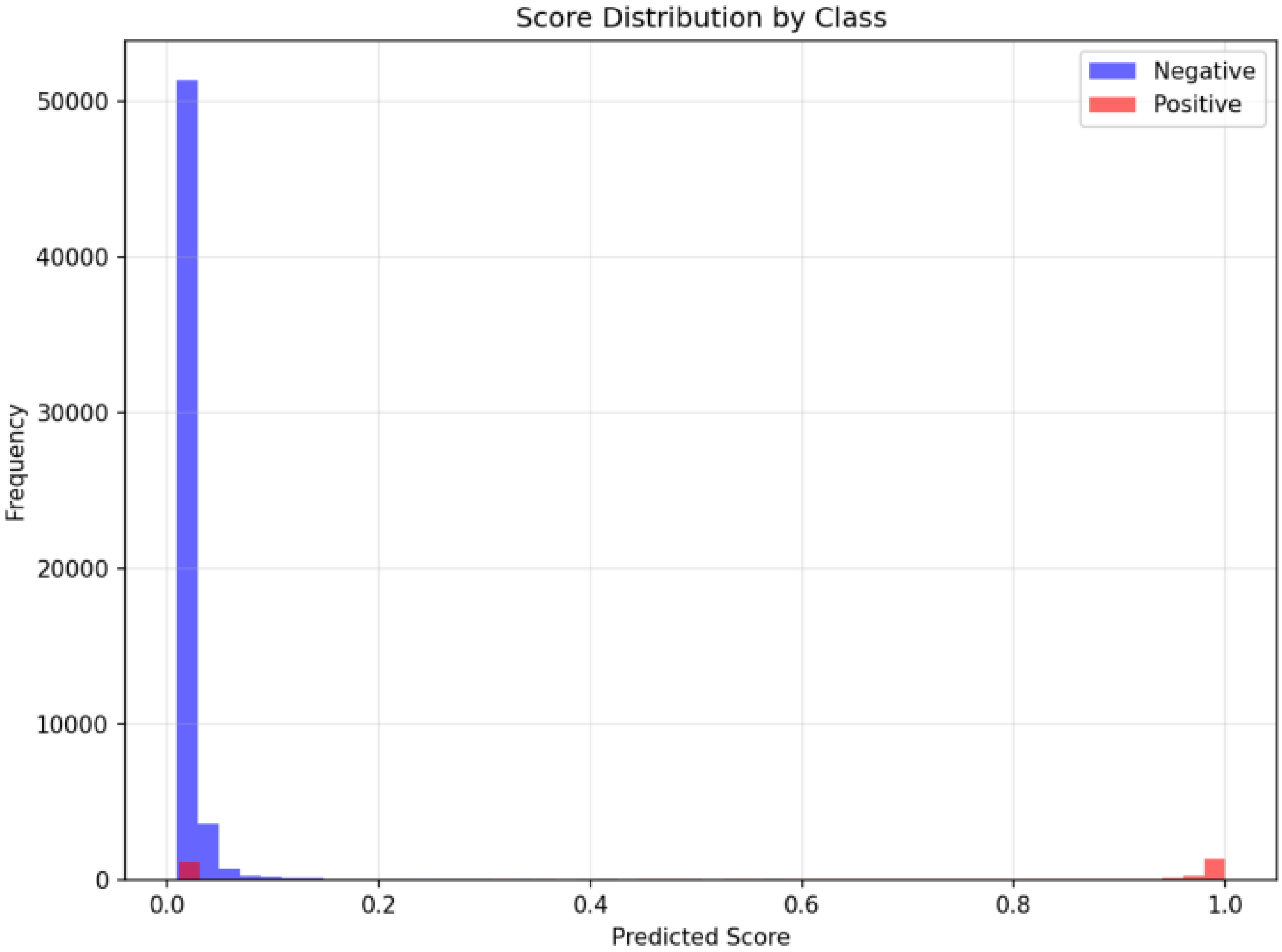
confusion_matrix_norm: Normalized rates: TPR=60.35%, TNR=99.87%.



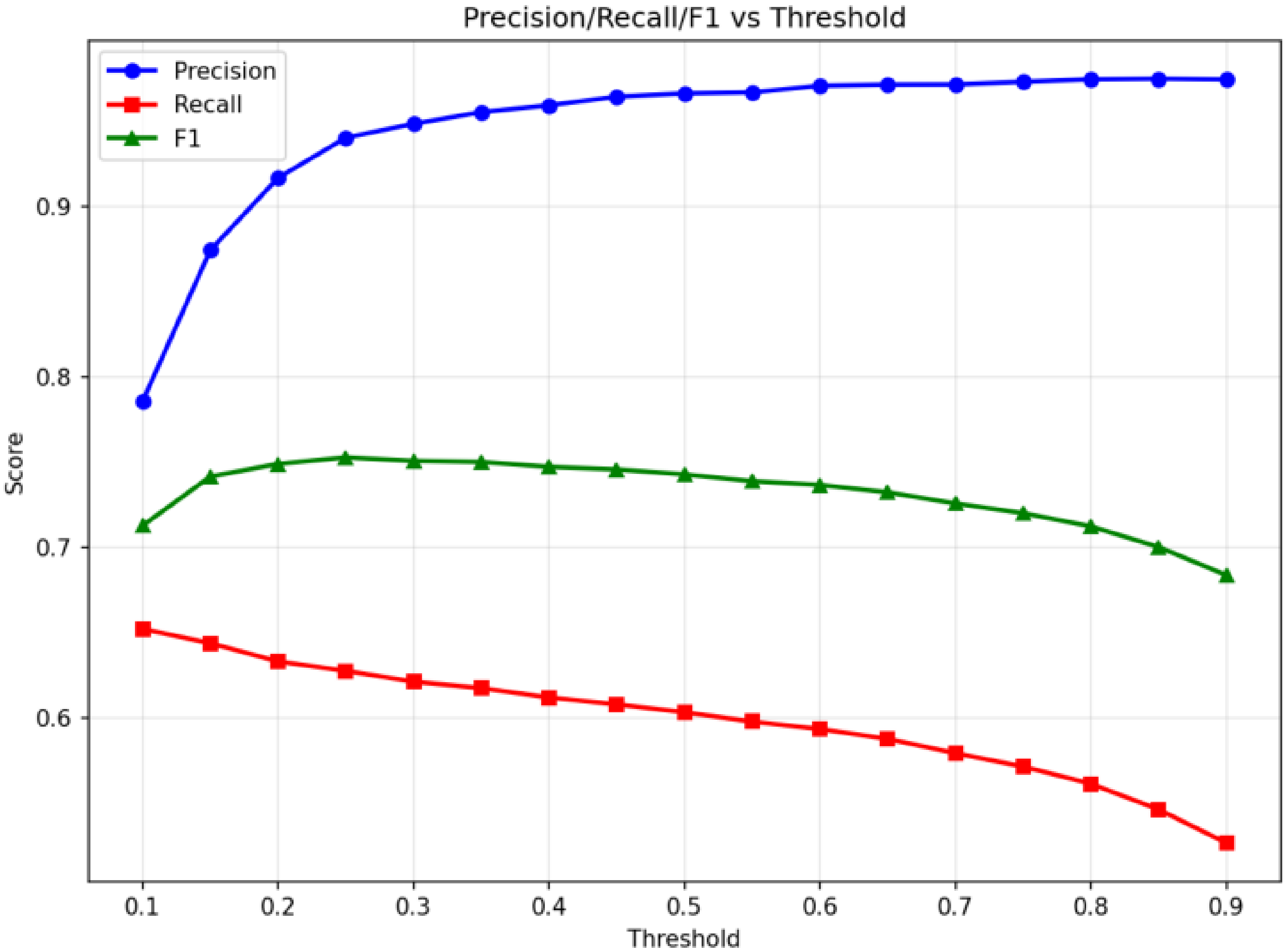
ks_curve: Maximum separation KS=0.6422 at threshold 0.12.



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



score_distribution: Mean score: positive=0.597, negative=0.024 (larger gap is better).



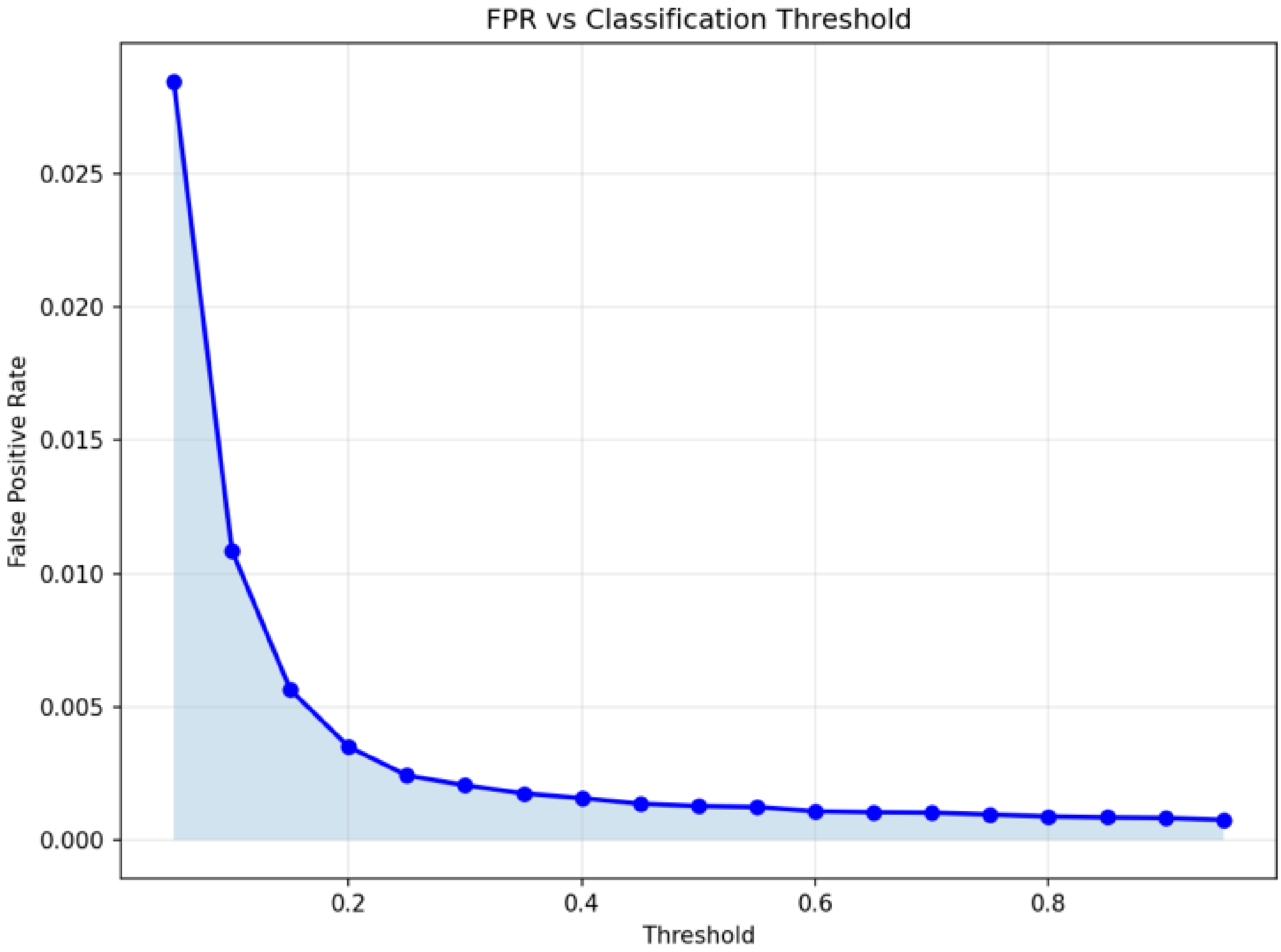
threshold_analysis: Best F1 in tested grid is 0.7530 at threshold 0.25.

2) Model Efficiency

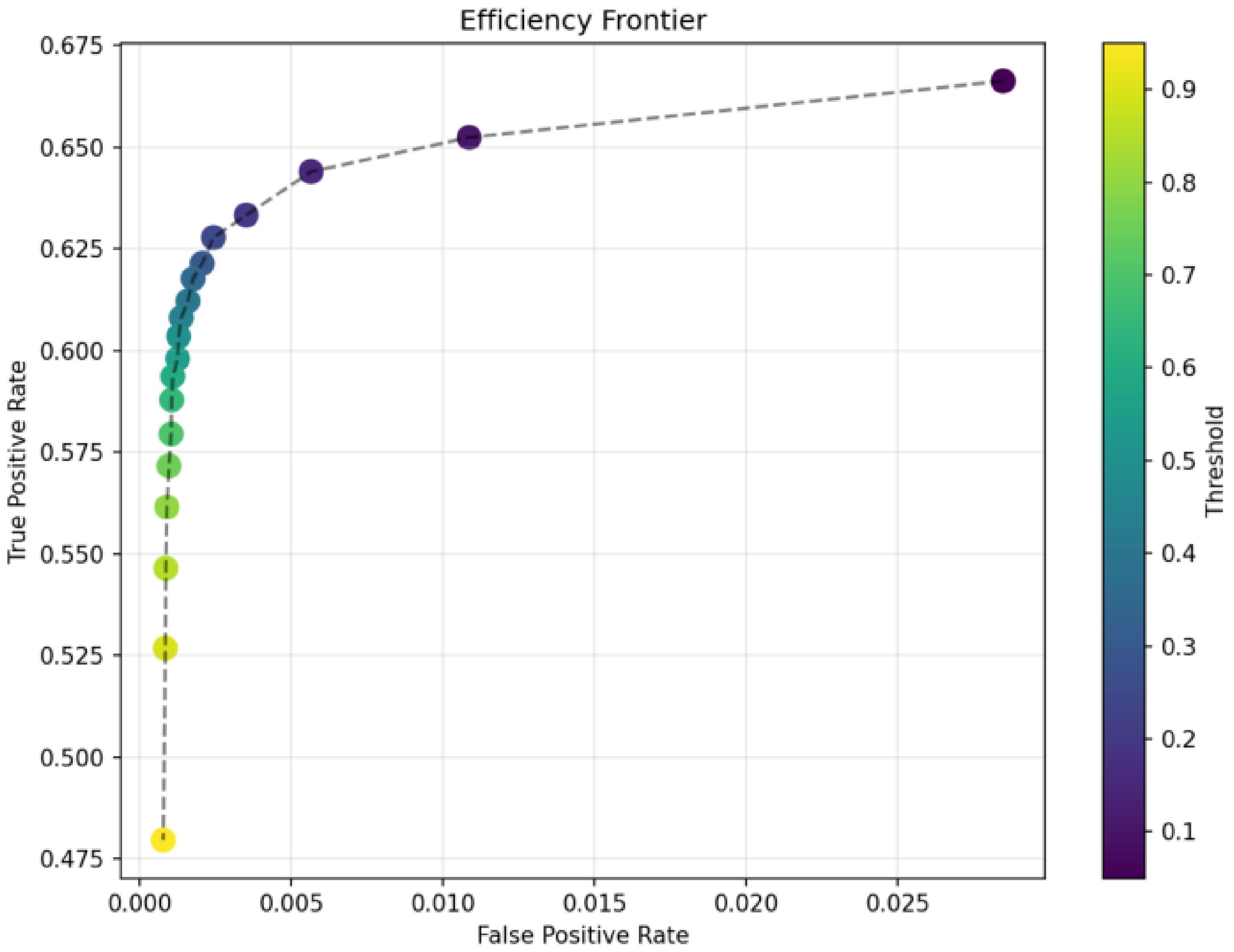
- fpr: 0.0013
- tn: 56473
- fp: 72
- threshold: 0.5000
- fpr_at_thresholds.t_0.05: 0.0285
- fpr_at_thresholds.t_0.10: 0.0109
- fpr_at_thresholds.t_0.15: 0.0056
- fpr_at_thresholds.t_0.20: 0.0035
- fpr_at_thresholds.t_0.25: 0.0024
- fpr_at_thresholds.t_0.30: 0.0021
- fpr_at_thresholds.t_0.35: 0.0018
- fpr_at_thresholds.t_0.40: 0.0016
- fpr_at_thresholds.t_0.45: 0.0014
- fpr_at_thresholds.t_0.50: 0.0013
- fpr_at_thresholds.t_0.55: 0.0012
- fpr_at_thresholds.t_0.60: 0.0011
- fpr_at_thresholds.t_0.65: 0.0010
- fpr_at_thresholds.t_0.70: 0.0010
- fpr_at_thresholds.t_0.75: 0.0010
- fpr_at_thresholds.t_0.80: 0.0009
- fpr_at_thresholds.t_0.85: 0.0008
- fpr_at_thresholds.t_0.90: 0.0008
- fpr_at_thresholds.t_0.95: 0.0008

2) Model Efficiency - Explanations

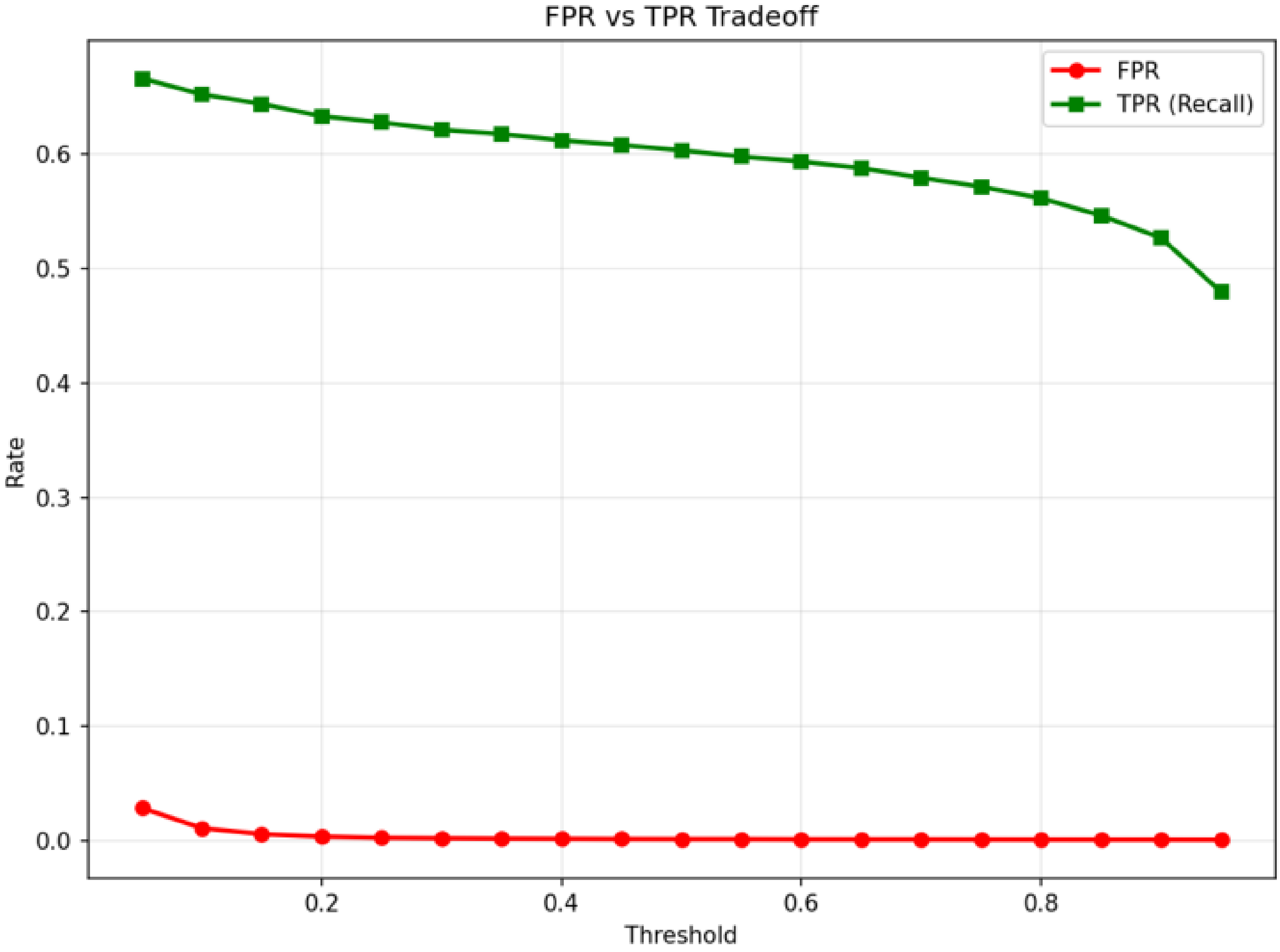
- At threshold 0.50, FPR=0.0013 (low); FP=72 out of 56545 negatives.
- At the same threshold, TPR=0.6035 with TP=2085 and FN=1370, showing the capture rate of positives.
- A threshold near 0.05 yields FPR≈0.0285 with TPR≈0.6663 if you want to target ~5% false positives.



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.



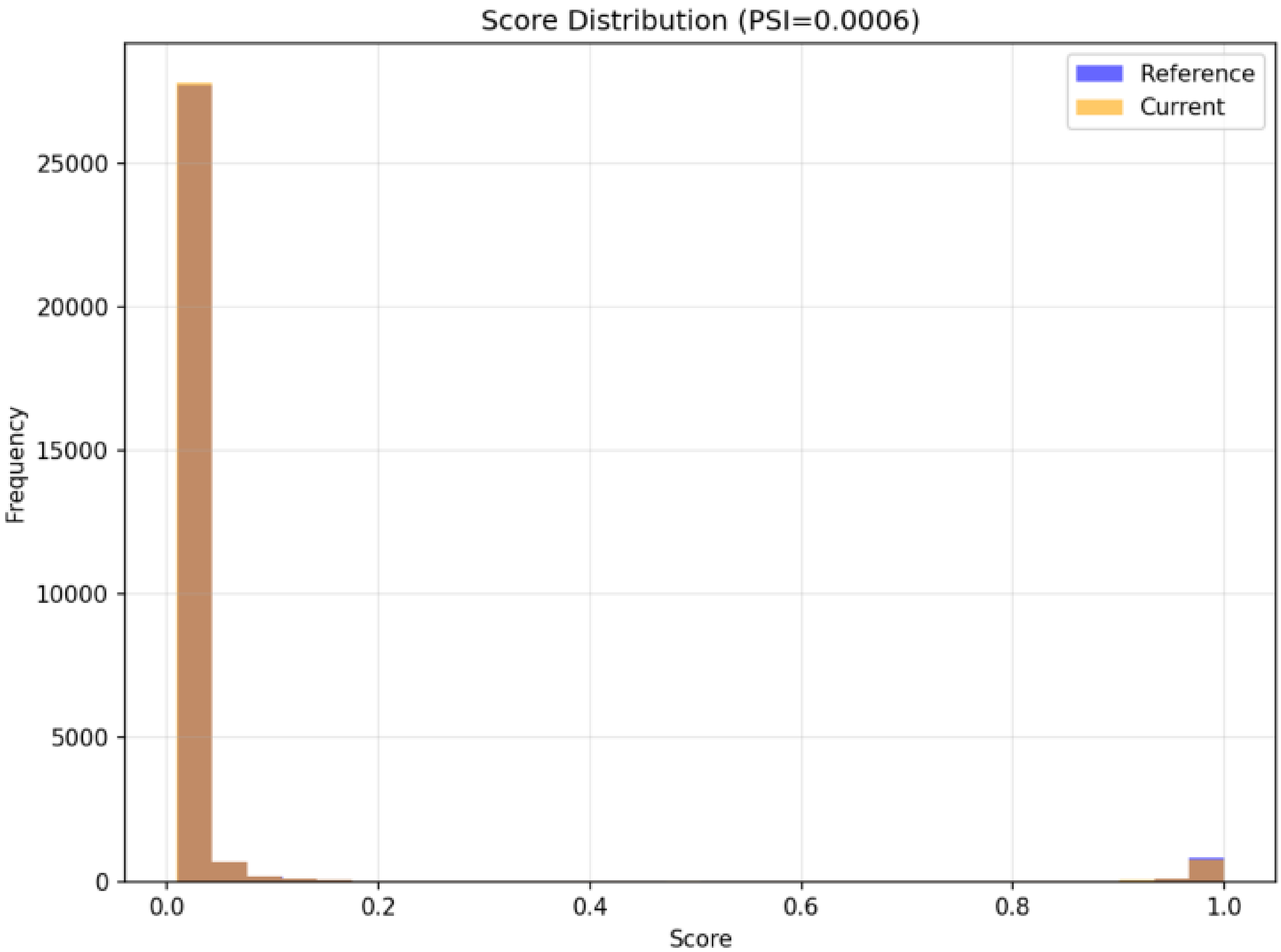
fpr_tpr_tradeoff: FPR and TPR curves highlight how recall drops as you reduce false positives.

3) Model Stability

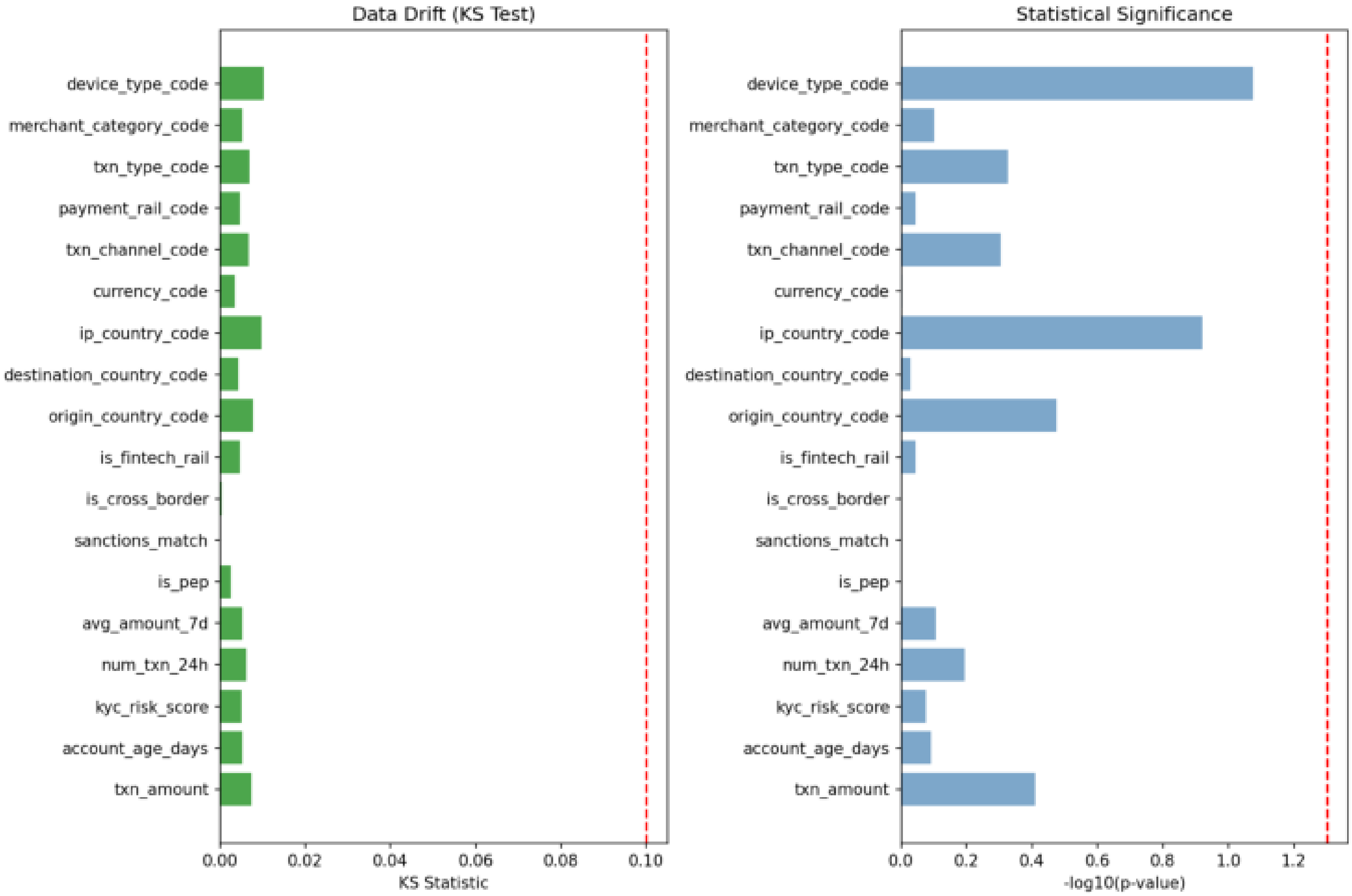
- psi: 0.0006
- cv_auc_roc_mean: 0.8298
- cv_auc_roc_std: 0.0107
- cv_auc_pr_mean: 0.6837
- cv_auc_pr_std: 0.0088
- bootstrap_auc_roc_mean: 0.8357
- bootstrap_auc_roc_ci_lower: 0.8238
- bootstrap_auc_roc_ci_upper: 0.8467
- concept_drift_detected: False
- concept_drift_score: 0.0225

3) Model Stability - Explanations

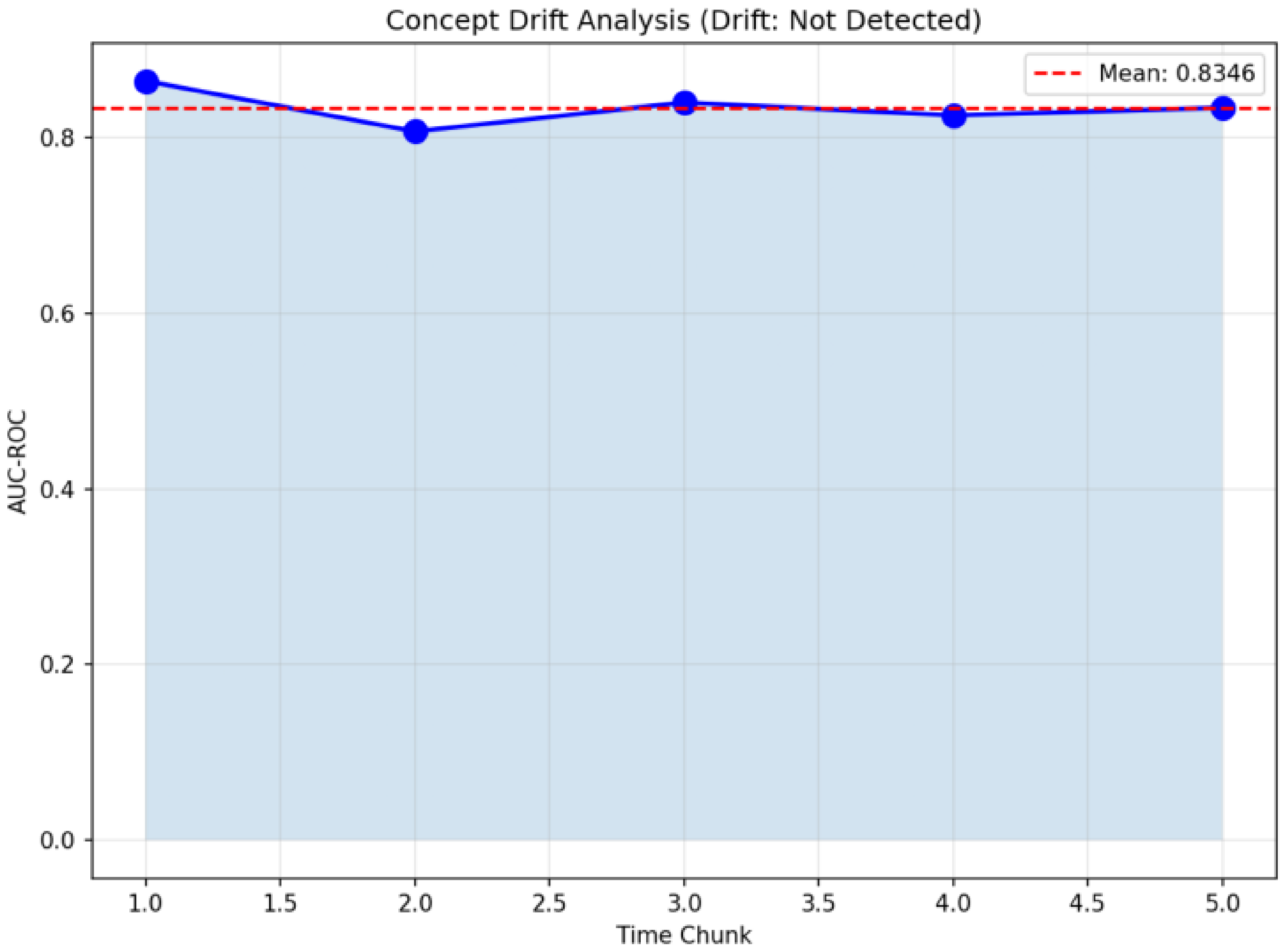
- PSI is 0.0006, indicating negligible score distribution shift between reference and current data.
- CV AUC-ROC is 0.8298 ± 0.0107 (stable across folds).
- CV AUC-PR is 0.6837 ± 0.0088 (stable across folds).
- Bootstrap AUC-ROC mean is 0.8357 with 95% CI [0.8238, 0.8467] (width 0.0230).
- No concept drift detected (score 0.0225), suggesting stable performance over time.
- Data drift flagged 0/18 numeric features; top drift signals: ['device_type_code', 'ip_country_code',
- 'origin_country_code'].



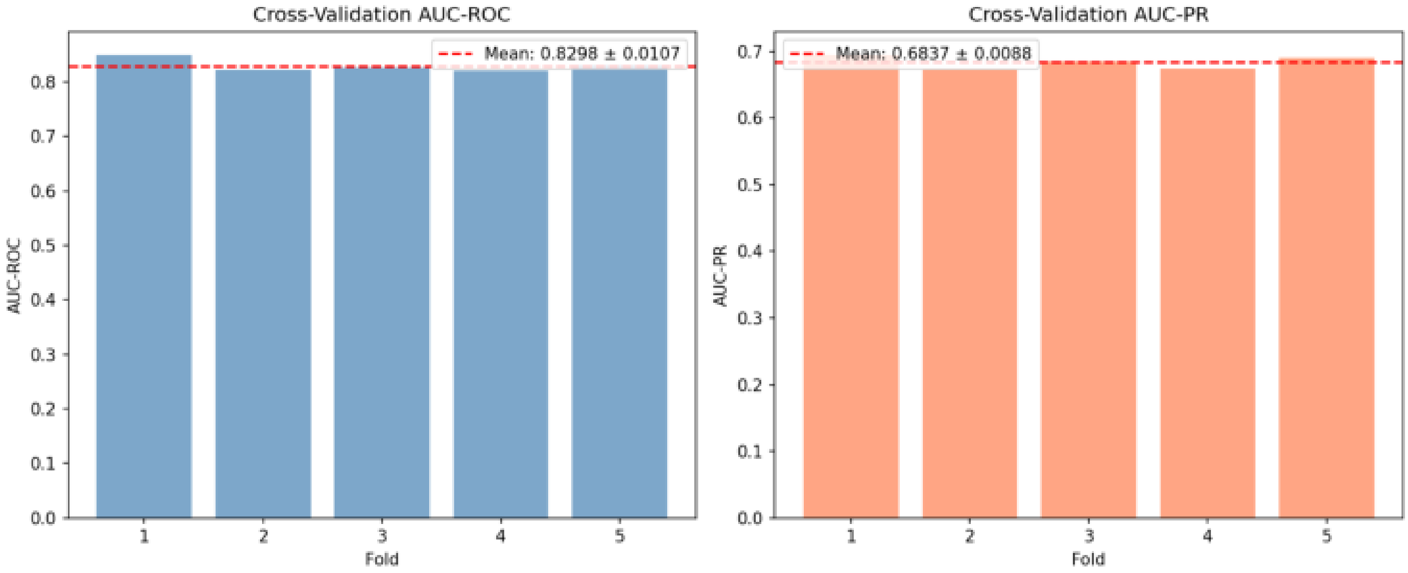
psi_distribution: Score distributions overlap consistent with PSI=0.0006.



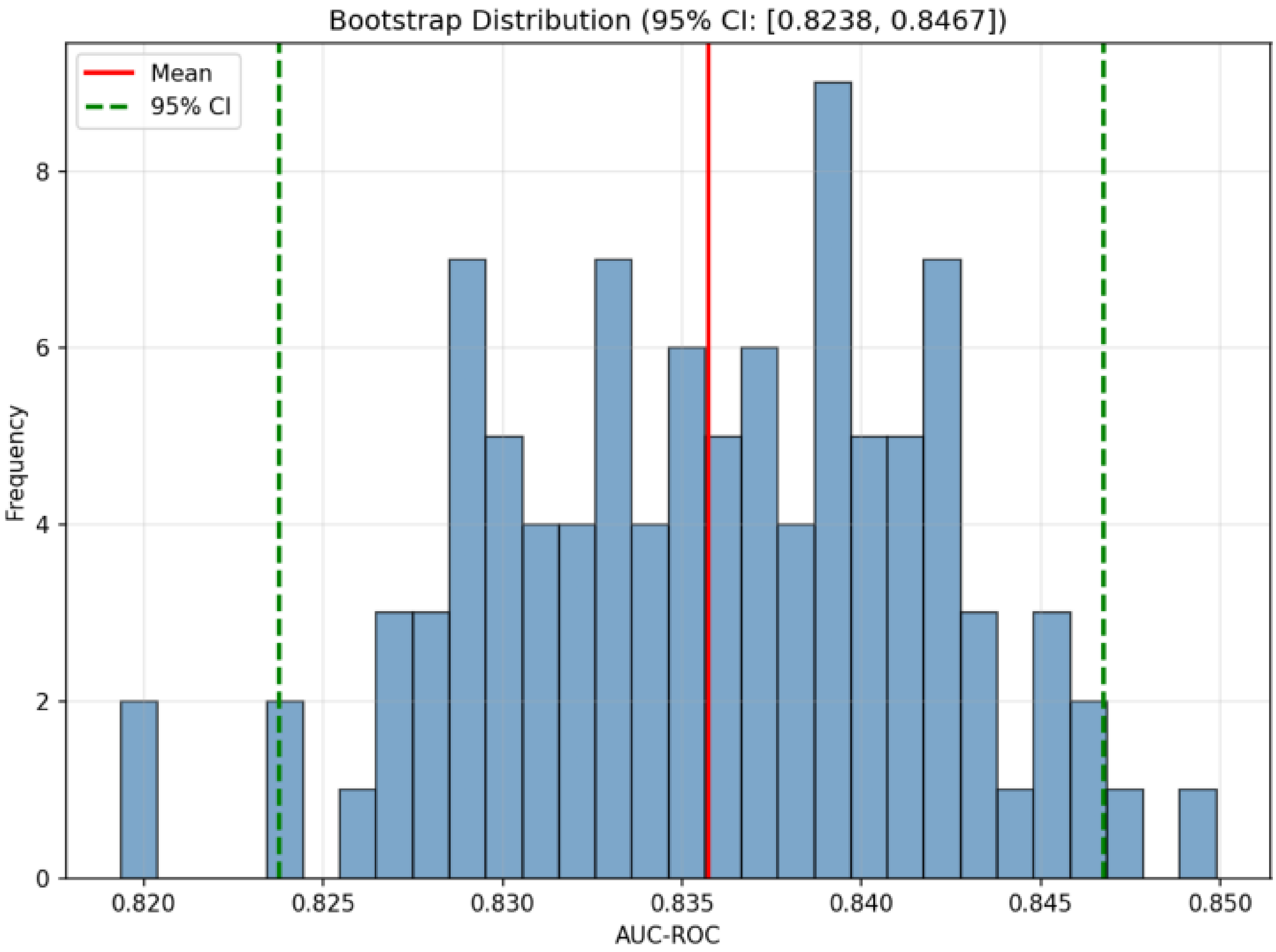
data_drift_heatmap: 0/18 features show drift; red bars highlight flagged features.



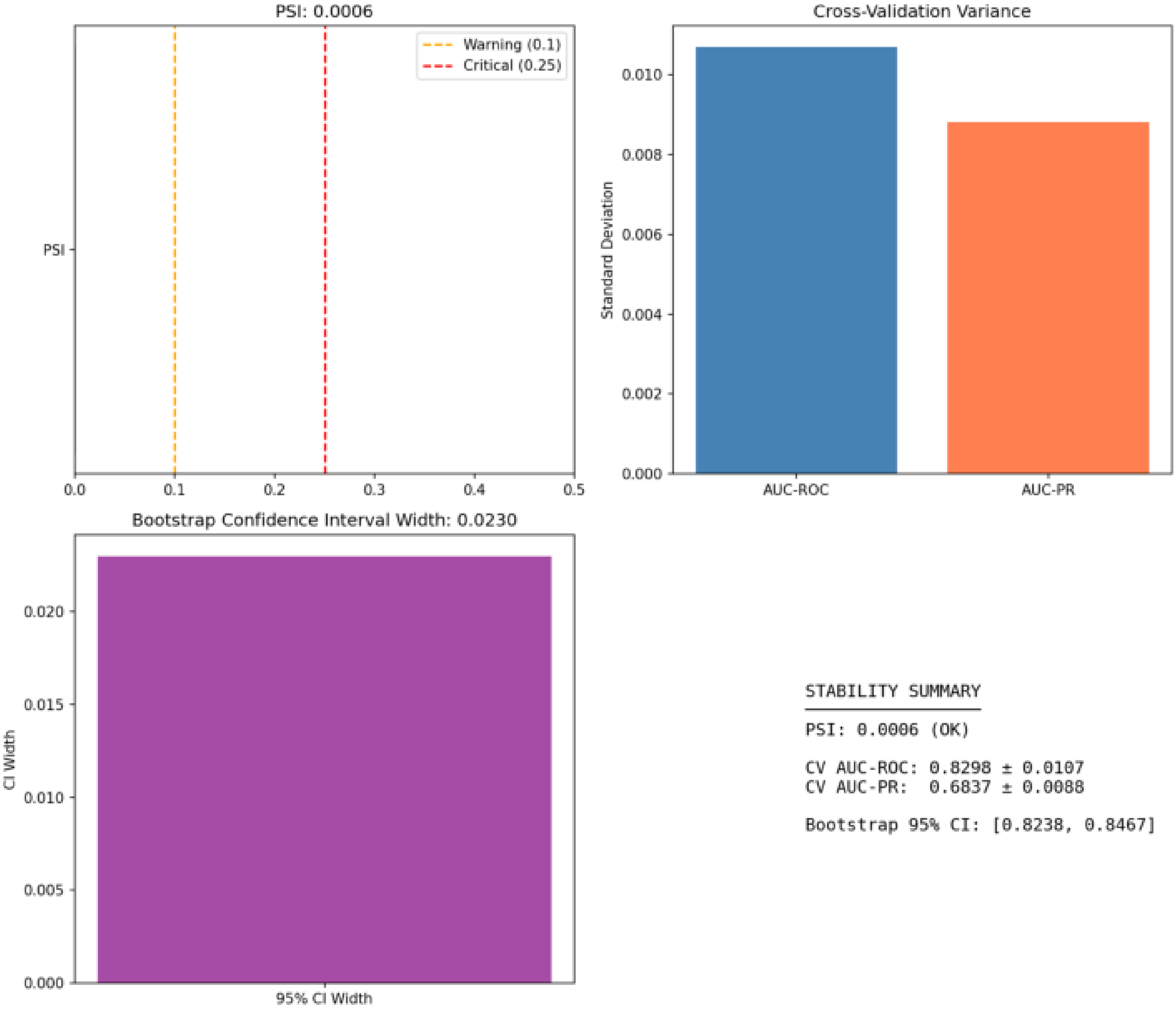
concept_drift: AUC across time chunks shows whether performance is stable over time.



cv_results: Fold scores cluster around ROC 0.8298 and PR 0.6837.



bootstrap_distribution: Bootstrap CI [0.8238, 0.8467] reflects performance stability.



```
STABILITY SUMMARY
PSI: 0.0006 (OK)

CV AUC-ROC: 0.8298 ± 0.0107
CV AUC-PR:  0.6837 ± 0.0088

Bootstrap 95% CI: [0.8238, 0.8467]
```

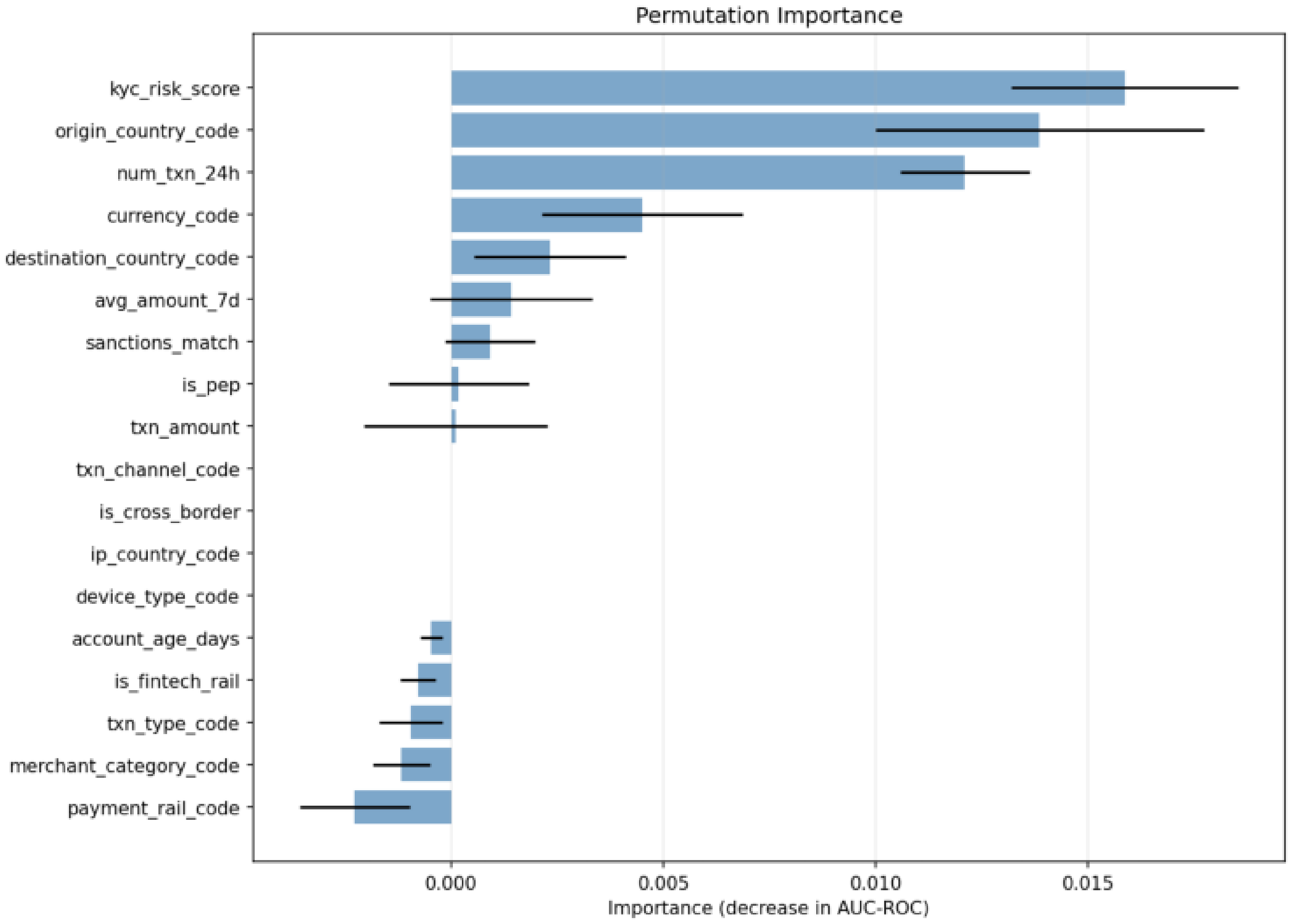
stability_summary: Dashboard summarizes PSI, CV variance, and bootstrap CI width.

4) Model Interpretability

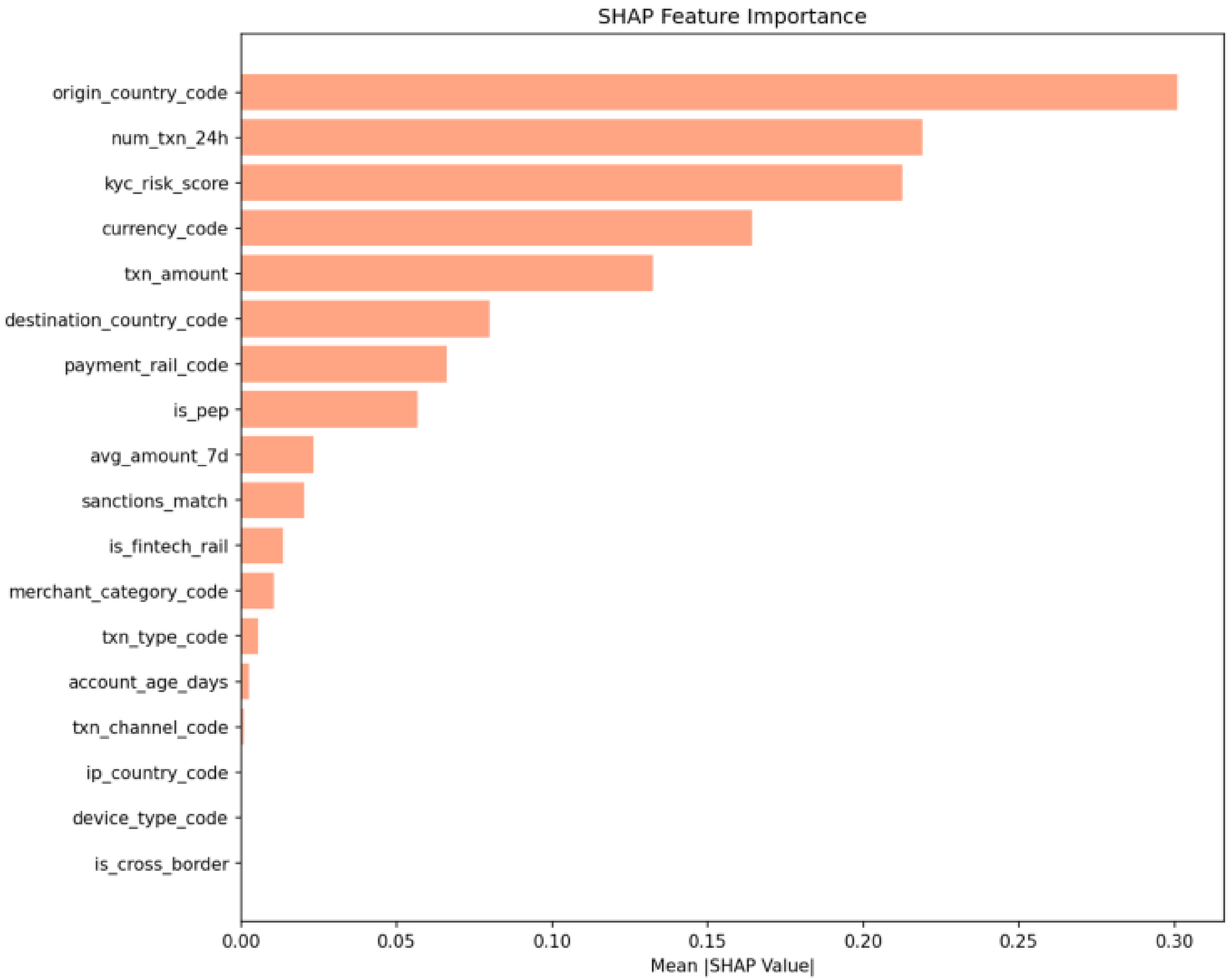
- model_type: tree
- methods_used: ['permutation', 'pdp', 'ice', 'shap', 'lime']
- perm_top_features: ['kyc_risk_score', 'origin_country_code', 'num_txn_24h', 'currency_code', 'destination_country_code']
- shap_top_features: ['origin_country_code', 'num_txn_24h', 'kyc_risk_score', 'currency_code', 'txn_amount']...
- lime_instances: 3
- 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

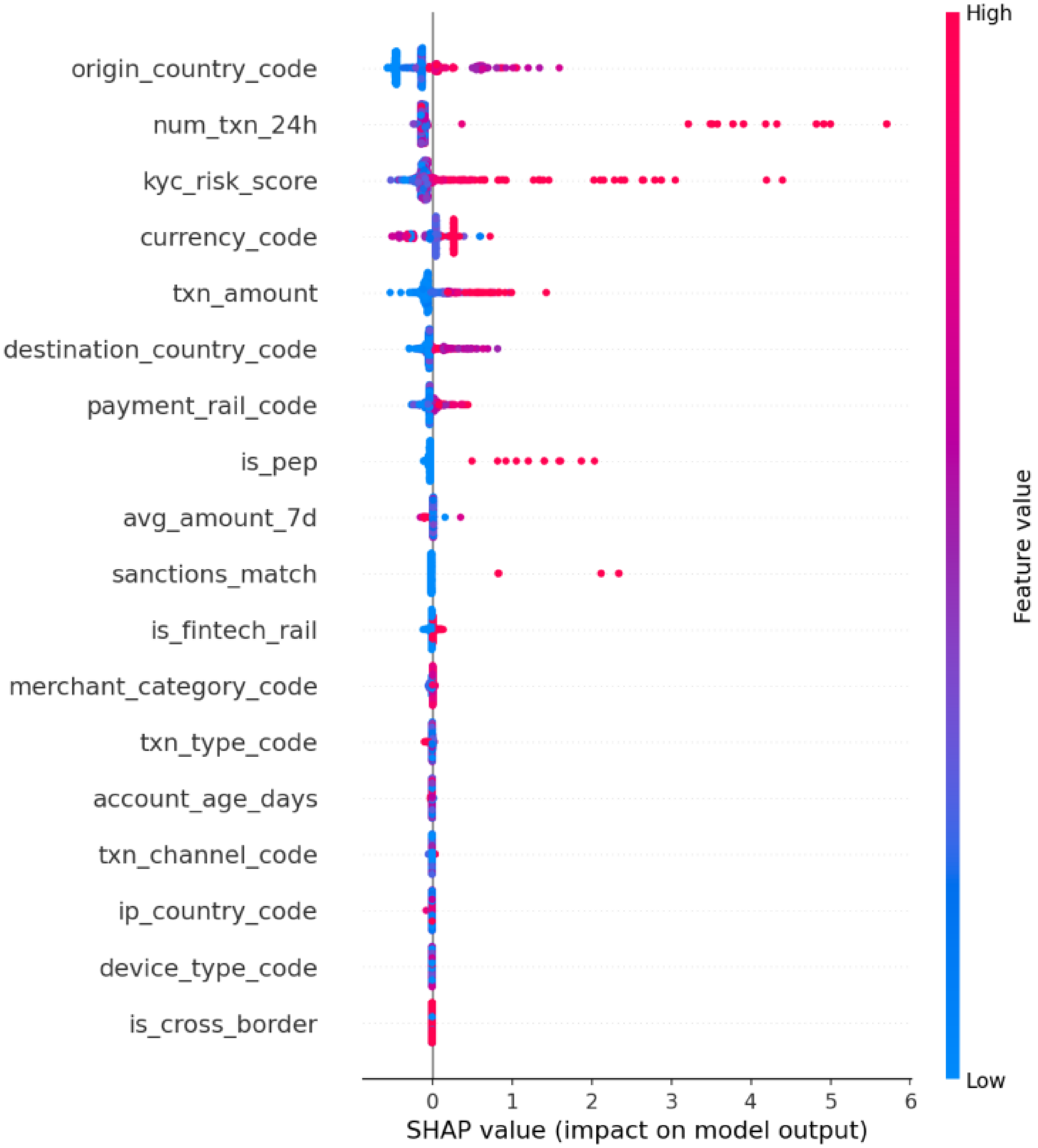
- Permutation importance highlights: ['kyc_risk_score', 'origin_country_code', 'num_txn_24h', 'currency_code',
- 'destination_country_code'].
- SHAP top features: ['origin_country_code', 'num_txn_24h', 'kyc_risk_score', 'currency_code', 'txn_amount'].
- LIME generated local explanations for 3 instances; top contributors for instance_6: ['num_txn_24h > 3.00',
- 'sanctions_match <= 0.00', 'is_pep <= 0.00', 'txn_amount > 333.44', 'kyc_risk_score > 46.42'].
- 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'].



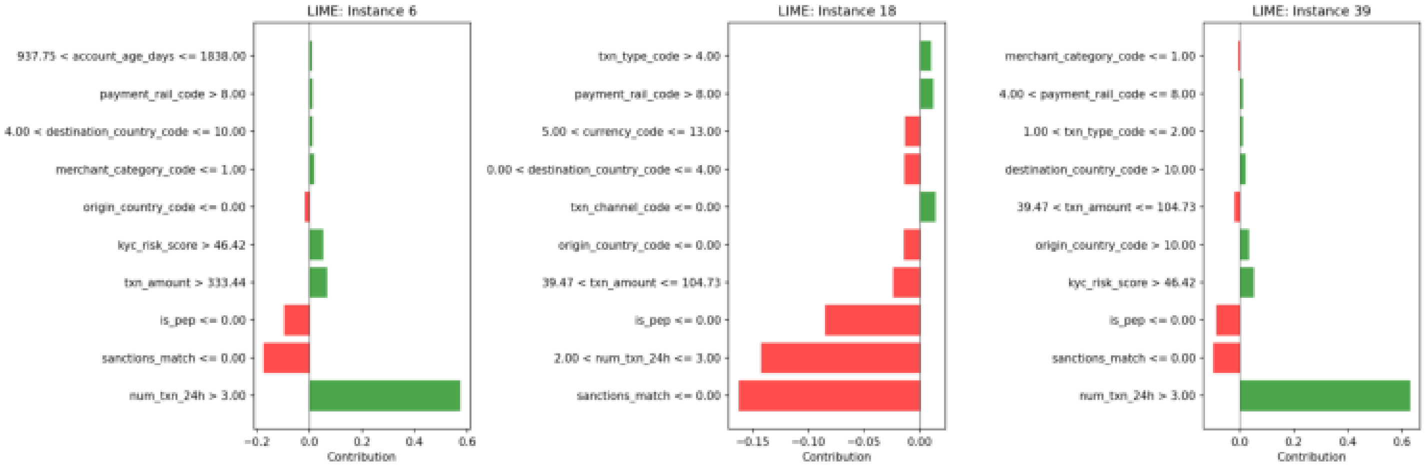
permutation_importance: Bars show how much AUC drops when each feature is permuted.



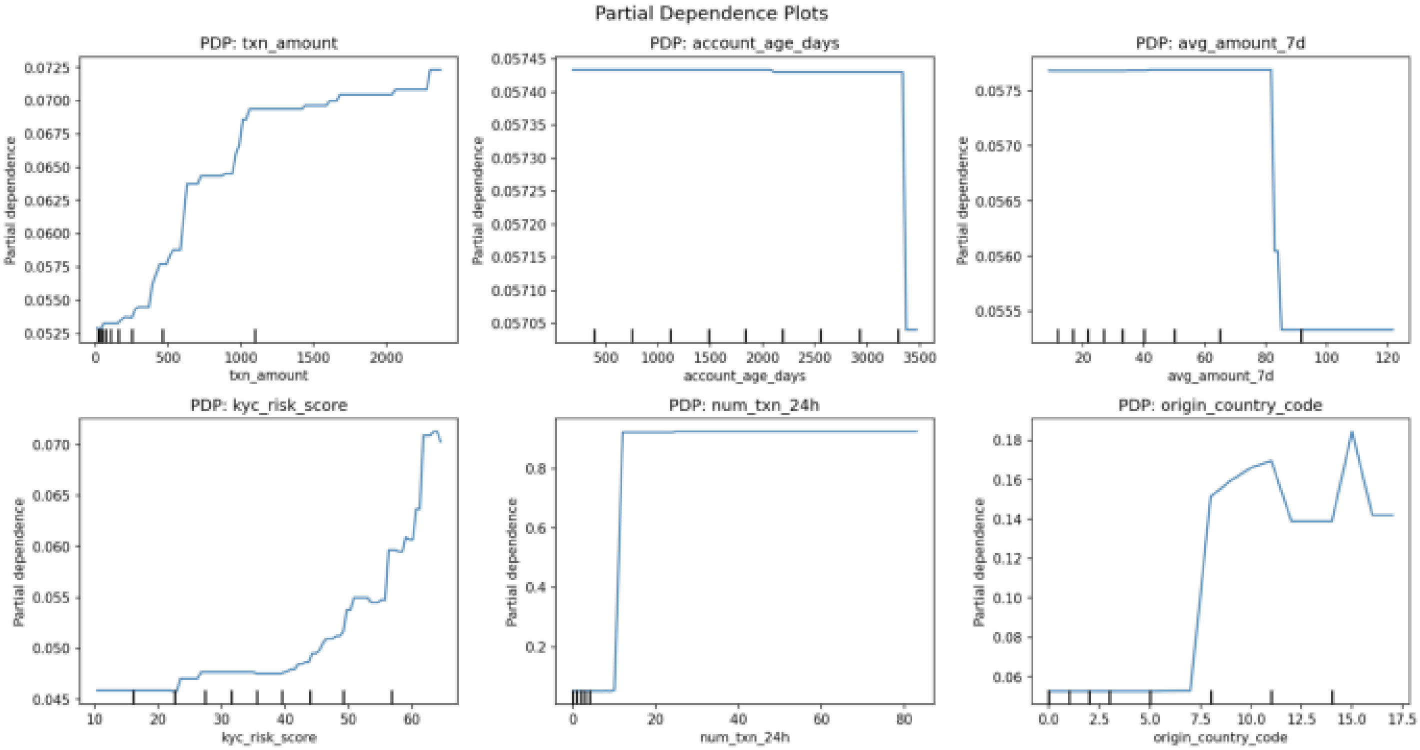
shap_bar: SHAP bar plot ranks global feature impact by mean absolute SHAP values.



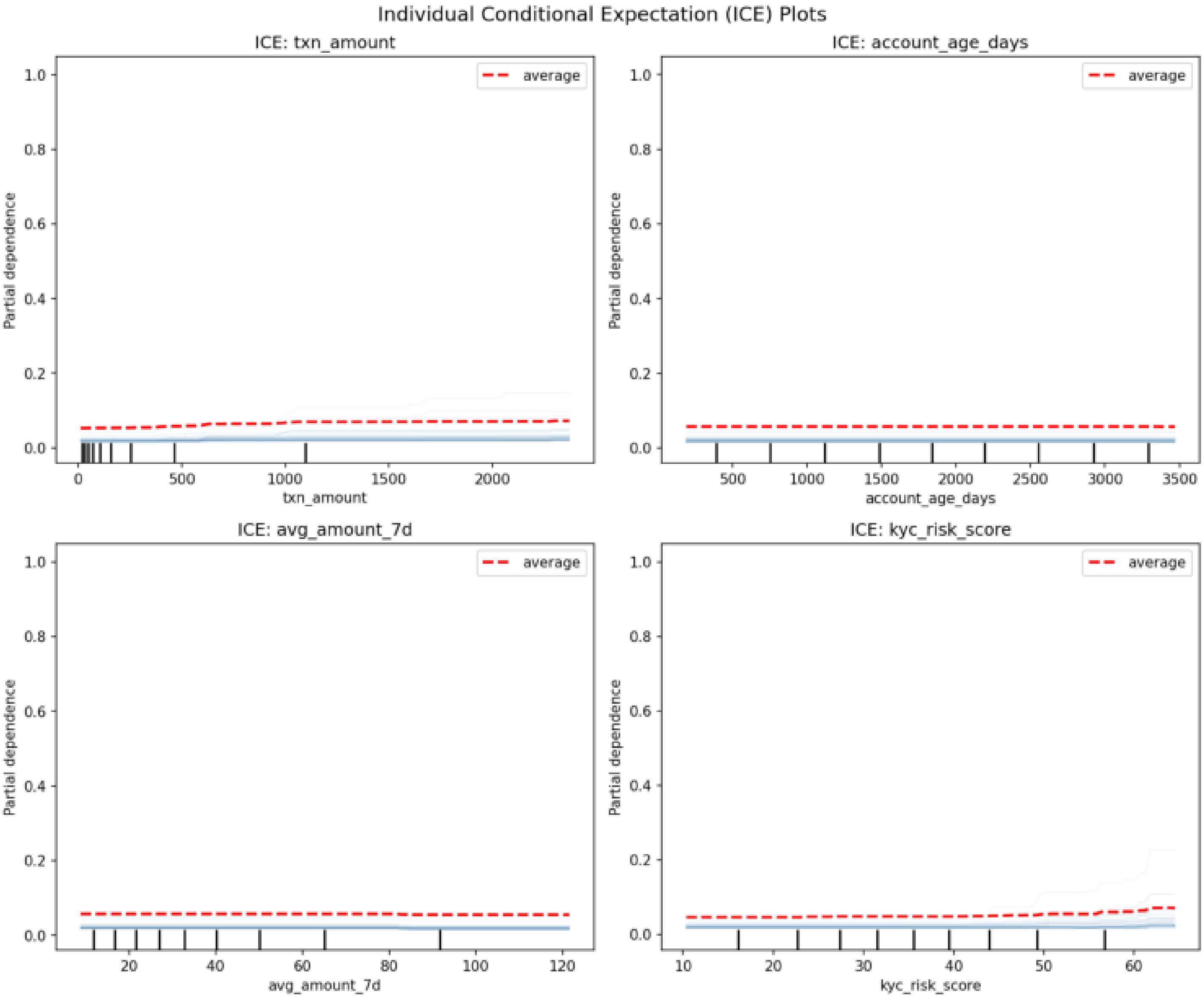
shap_beeswarm: SHAP beeswarm shows both impact size and direction per feature.



lime_explanation: LIME bars show local feature contributions for selected instances.



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



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