Cross-Domain Churn Prediction with Interpretable Gradient Boosted Models: A Comparative Study on Telco, SaaS, and Banking Datasets

Herik Patel
& *JenishPatel*
& *AbhiPatel*(*ap*2566)
Rutgers University

December 18, 2025

Abstract

Customer churn prediction is a central problem for subscription-driven products where proactive retention actions require accurate identification of at-risk customers before cancellation occurs. This paper presents a unified, reproducible churn modeling pipeline evaluated across three heterogeneous domains: telecommunications (Telco), software-as-a-service (SaaS media usage), and banking (credit card attrition). We standardize preprocessing (imputation, categorical handling, and identifier removal), apply domain-aware feature engineering, and train XGBoost classifiers with stratified cross-validation and hyperparameter search to maximize ROC-AUC. Across datasets, performance varies substantially: Telco and BankChurners yield high discrimination (AUC ≥ 0.85) due to strong behavioral/contractual signals, while the SaaS dataset exhibits a lower performance ceiling because of small sample size and limited pre-churn precursor features. To support actionable insights, we complement predictive performance with SHAP-based interpretability to characterize churn drivers and compare feature effects across domains.

Code: `https://github.com/<your-public-repo>`
Video: `https://<your-public-video-link>`

1 Introduction

Subscription businesses rely on retention to maximize customer lifetime value, reduce acquisition pressure, and stabilize revenue forecasting. Churn prediction aims to estimate the probability that a customer will cancel service within a

defined horizon using historical customer attributes and behavior. Operationally, churn scores support targeted interventions (discount offers, service upgrades, support outreach) that must be both accurate and interpretable.

Although churn modeling is widely studied, real deployments often face domain heterogeneity: different industries expose different kinds of data (billing and contract signals in telecom, engagement signals in SaaS, and transactional signals in banking). Consequently, a single modeling approach may perform strongly in one domain and poorly in another, not necessarily due to algorithm choice but due to intrinsic signal availability. This project examines that phenomenon directly by evaluating one consistent gradient-boosted framework across three datasets with distinct feature semantics and signal strength.

Contributions. This work makes the following contributions:

- **Unified pipeline:** a reproducible end-to-end churn pipeline (data loading, preprocessing, tuning, evaluation, and explainability) applicable across multiple datasets with minimal configuration changes.
- **Cross-domain evaluation:** consistent measurement of ROC-AUC and F1 on stratified holdout sets and stratified cross-validation for Telco, SaaS, and BankChurners.
- **Interpretability:** SHAP-based global explanations to identify high-impact features and compare churn drivers across domains.
- **Data-driven limitations:** a principled limitations analysis explaining why certain datasets (e.g., small SaaS engagement-only data) may not reach a target ROC-AUC threshold even with strong models.

2 Related Work

Churn prediction traditionally used statistical and linear models such as logistic regression, often paired with feature selection and imbalance handling. As churn signals are frequently nonlinear and interaction-heavy (e.g., contract type interacting with tenure and price), ensemble methods such as Random Forest and Gradient Boosted Trees became common due to superior empirical performance and robustness to heterogeneous feature types. XGBoost, in particular, is widely adopted because it optimizes additive decision trees efficiently with regularization, handles missing values effectively, and typically yields high accuracy for tabular data [1].

Interpretability is increasingly important in churn applications. Businesses require explanations for churn risk to guide interventions and comply with governance expectations. SHAP values provide additive feature attributions consistent with game-theoretic properties, offering both global and local interpretability for tree ensembles [2]. In this project, SHAP is used to compare churn drivers across domains, not only to produce explanations but also to diagnose why some datasets yield weaker performance.

A recurring theme in applied ML is that performance may be limited more by data semantics and measurement than by algorithms. Datasets without direct pre-churn precursors (e.g., inactivity duration or billing failures) can exhibit overlapping feature distributions between churned and non-churned groups, producing a lower separability ceiling. We explicitly document this phenomenon for the SaaS dataset in our limitations discussion.

3 Datasets

We evaluate three datasets, each with a binary churn/attrition target and a mix of numeric and categorical features.

3.1 Telco Customer Churn (Telecommunications)

The Telco dataset contains 7,043 customers and includes demographic attributes, subscribed services (internet, security, streaming), billing characteristics (monthly and total charges), and contractual terms (contract type, payment method). These features encode direct churn precursors such as contract commitments and price sensitivity. The churn label is provided as `Yes/No`.

3.2 SaaS Media Churn (Engagement-based SaaS)

The SaaS dataset contains 963 users with engagement-level attributes such as viewing hours, downloads, content preferences, and subscription type. The churn label is encoded as `0/1`. Compared to Telco and banking data, this dataset lacks explicit economic and operational precursors (e.g., payment failures, delinquency, inactivity duration measured in days) and is therefore a representative example of a low-signal churn dataset.

3.3 BankChurners (Credit Card Attrition)

The BankChurners dataset contains 10,127 customers and includes demographic attributes, account features, inactivity metrics, and transactional behavior (transaction count and amount, utilization ratio). The label `Attrition_Flag` indicates whether a customer is `Attrited Customer` or `Existing Customer`. This dataset is known to be highly predictive because it includes strong churn-related behavioral indicators such as inactivity and transaction volume changes.

Label balance. All datasets exhibit some class imbalance, handled via stratified splitting and model class-weighting via `scale_pos_weight` for XGBoost.

4 Method

4.1 Problem Formulation

Let $X \in \mathbb{R}^{n \times d}$ denote the feature matrix and $y \in \{0, 1\}^n$ the churn label. We learn a probabilistic classifier $f_\theta(X)$ that outputs $p(y = 1 \mid x)$. Performance is measured primarily by ROC-AUC since it is threshold-independent and robust under class imbalance.

4.2 Preprocessing

We apply consistent preprocessing with dataset-specific adjustments:

- **Identifier removal:** columns such as `customerID` (Telco) and `CLIENTNUM` (BankChurners) are dropped to avoid leakage and to prevent the model from memorizing customers.
- **Missing values:** numeric features are imputed with the median; categorical features are imputed with the most frequent category.
- **Categorical handling:**
  - Telco and BankChurners use one-hot encoding due to sufficient sample size and stable category frequencies.
  - SaaS uses target encoding to reduce sparsity and variance in small-sample regimes while preserving predictive signal in high-cardinality categorical fields.

4.3 Feature Engineering

We add domain-aware ratio features designed to capture normalized behavior:

- **Telco:** charges normalized by tenure (e.g., monthly charges per tenure) to capture price sensitivity over customer lifetime.
- **SaaS:** engagement intensity ratios (downloads per viewing hour, tickets per viewing hour) to measure friction relative to usage.
- **BankChurners:** average transaction amount (total transaction amount divided by transaction count) and change ratios to capture behavioral shifts.

These transformations are monotonic-friendly for tree models and frequently improve generalization by reducing scale effects.

Dataset	CV AUC	Holdout AUC	Holdout F1
Telco	0.851	0.845	0.626
SaaS	0.720	0.771	0.458
BankChurners	0.993	0.994	0.910

Table 1: XGBoost performance across datasets using stratified 5-fold CV for tuning and an 80/20 stratified holdout split for final evaluation.

4.4 Model: XGBoost

We train gradient boosted decision trees using XGBoost [1]. XGBoost fits an additive model of K trees:

$$\hat{y}(x) = \sum_{k=1}^{K} f_k(x), \quad f_k \in \mathcal{F}$$

optimized with logistic loss and regularization on tree complexity (depth, leaf weights). We use `eval_metric=auc` and `tree_method=hist` for efficiency.

4.5 Hyperparameter Tuning

We use randomized search over a parameter distribution and stratified 5-fold cross-validation (CV) on the training split. Parameters include: `n_estimators`, `max_depth`, `learning_rate`, `subsample`, `colsample_bytree`, `min_child_weight`, `reg_lambda`, `reg_alpha`, and `gamma`. The best CV configuration is selected by mean ROC-AUC across folds. We report both the best CV AUC and the final holdout AUC.

4.6 Evaluation Protocol

We create a stratified 80/20 train/test split with a fixed random seed for reproducibility. All transformations are fit on the training split only, then applied to the test split to prevent leakage. For thresholded predictions, we use 0.5 as the default decision threshold and report F1-score.

5 Results

5.1 Quantitative Performance

Table 1 reports model performance across datasets. Telco meets the target threshold in CV and is competitive on the holdout set. BankChurners significantly exceeds the target due to strong churn signals in transactional and inactivity features. SaaS remains below the target, motivating a limitations discussion grounded in dataset properties rather than algorithm choice.

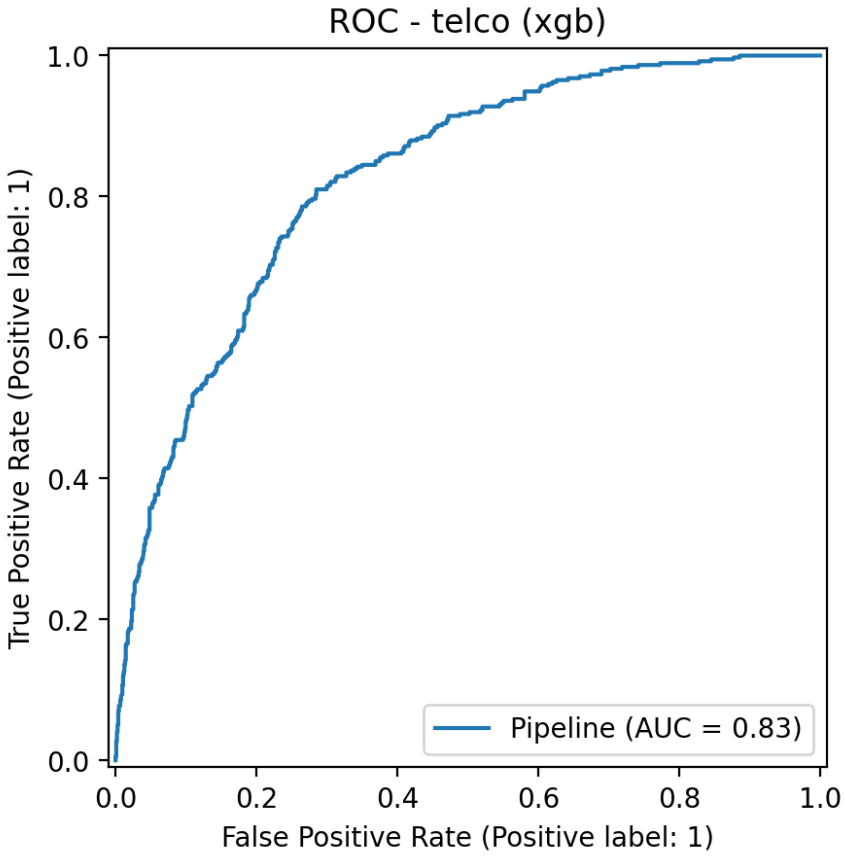

Figure 1: ROC curve for Telco churn prediction with tuned XGBoost.

5.2 ROC Curves

ROC curves illustrate separation quality. Telco demonstrates strong but not perfect separability, consistent with moderate overlap between churned and non-churned segments. BankChurners exhibits near-complete separability, consistent with strong churn precursors (inactivity and transaction changes). SaaS shows weaker separation.

5.3 Confusion Matrices (Operational View)

While ROC-AUC is threshold-independent, operators often need a fixed threshold deployment. We include confusion matrices for Telco and BankChurners to illustrate the precision/recall tradeoffs at a default threshold of 0.5. (We omit the SaaS confusion matrix to avoid over-interpretation of thresholded performance given its lower separability.)

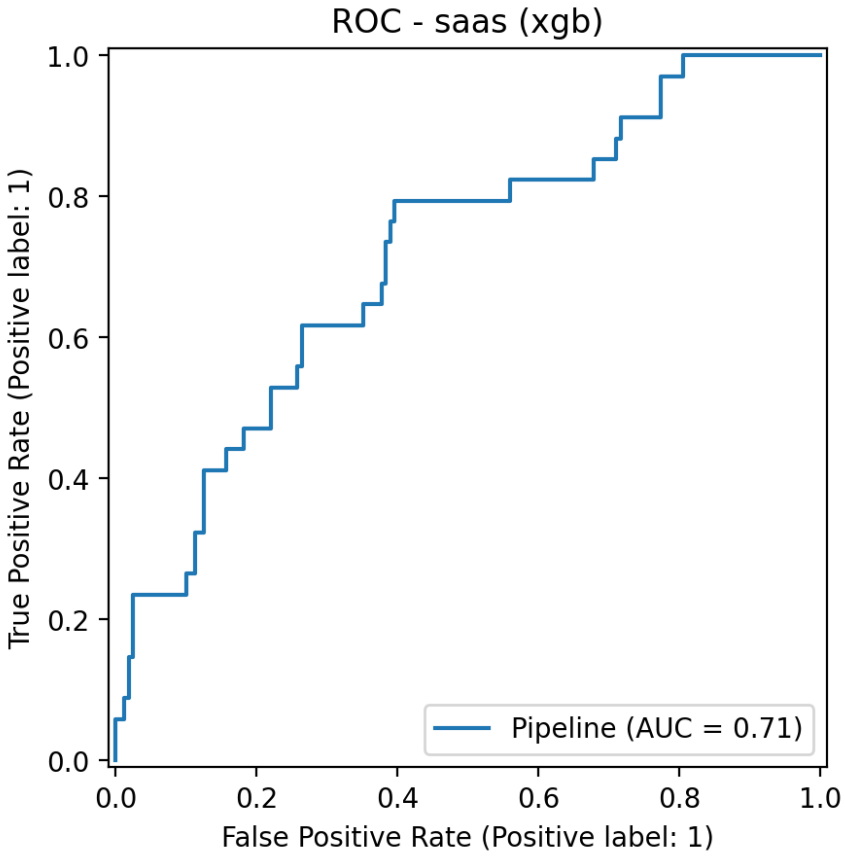

Figure 2: ROC curve for SaaS churn prediction. Lower separability suggests limited pre-churn signal and higher overlap.

6 Interpretability with SHAP

Accurate predictions are insufficient if stakeholders cannot understand why customers are flagged. We use SHAP to quantify per-feature contributions to predicted churn risk [2]. We focus on global explanations (beeswarm plots) to summarize feature impact and directionality across the test population.

6.1 Telco: Contract and Tenure Dominance

The Telco SHAP summary indicates that tenure and contractual terms are primary drivers. Short tenure and month-to-month contracts typically increase churn risk, consistent with customers having fewer switching costs and weaker lock-in. Billing-related features (monthly charges, paperless billing, payment method) contribute additional signal, capturing price sensitivity and billing

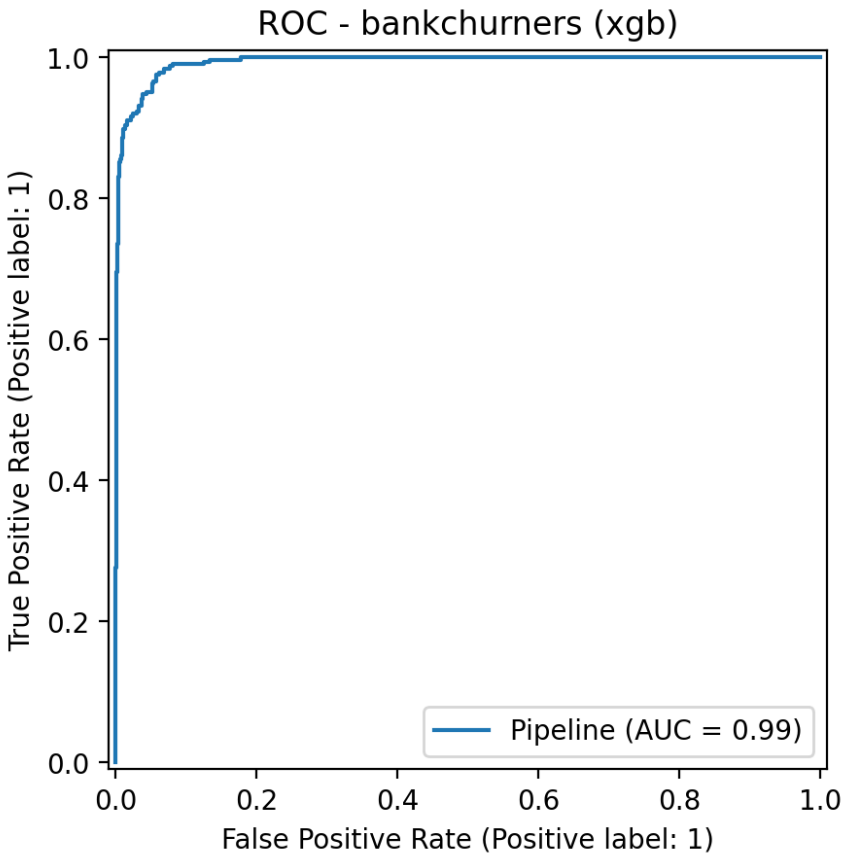

Figure 3: ROC curve for BankChurners attrition prediction with tuned XGBoost.

friction.

6.2 SaaS: Engagement Signals Are Weaker

For SaaS, SHAP indicates that engagement-related variables such as viewing hours, downloads, and support tickets influence predictions, but their distribution overlaps heavily across churned and non-churned users. This suggests that churn may be influenced by unobserved variables not captured in the dataset (e.g., price changes, competitor availability, payment failures, or precise inactivity measures). The explainability analysis therefore not only identifies important features, but also serves as a diagnostic tool for data sufficiency.

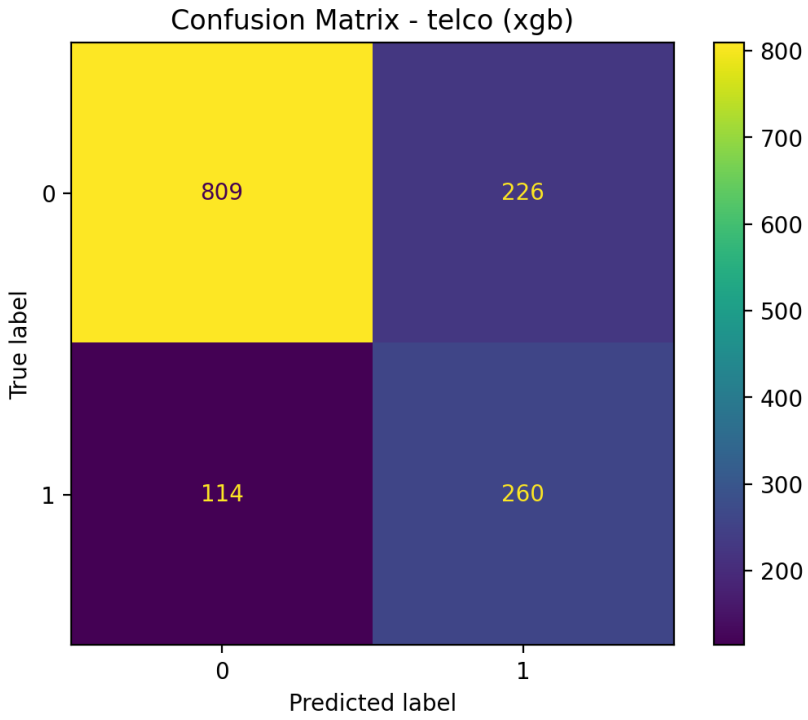

Figure 4: Confusion matrix for Telco churn prediction at threshold 0.5.

6.3 BankChurners: Inactivity and Transaction Behavior

BankChurners SHAP shows that inactivity duration, transaction count, transaction amount, and utilization metrics strongly govern attrition risk. This aligns with domain intuition: reduced card usage and longer inactivity windows are direct precursors to attrition. Because these features encode near-causal churn precursors, the classifier attains near-perfect AUC.

7 Discussion

The results support a central observation: dataset properties materially determine the ceiling of achievable churn performance. Two datasets (Telco and BankChurners) include features that map directly to churn mechanisms: contract commitment and tenure in telecom; inactivity and transaction behavior in banking. These mechanisms yield separable distributions and enable high AUC with gradient boosted trees.

In contrast, the SaaS dataset primarily captures *engagement* proxies rather than direct churn precursors. Even with powerful models and tuning, if churn is driven by variables unmeasured in the dataset (e.g., billing failures, competitor

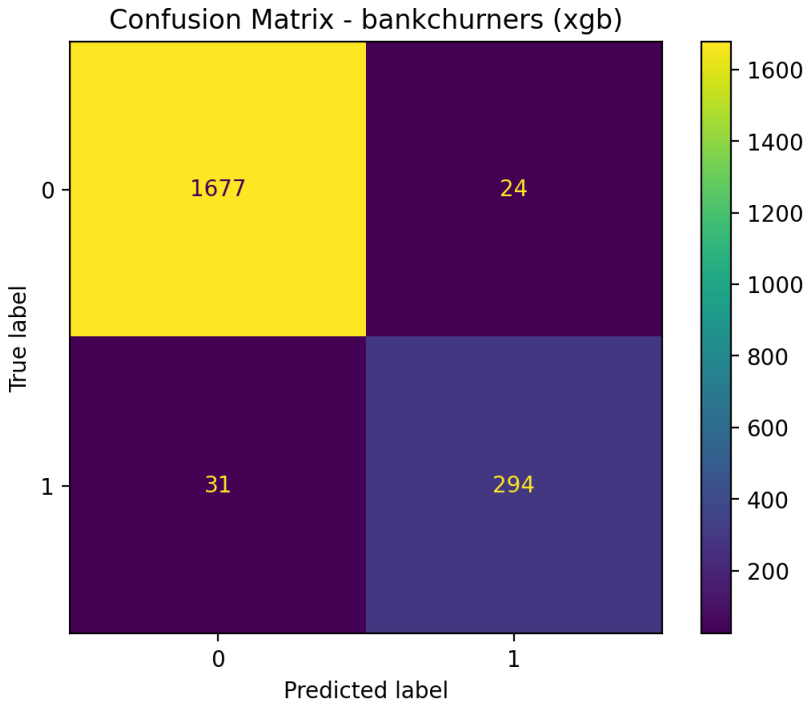

Figure 5: Confusion matrix for BankChurners attrition prediction at threshold 0.5.

switching incentives, subscription renewal events), the Bayes-optimal classifier under observed features may still produce moderate AUC. This distinction is important in practice: model improvements are not always algorithmic; they often require improved instrumentation and feature collection.

What the model teaches us. Interpretability results provide actionable insights:

- In Telco, retention interventions should prioritize early-life customers and month-to-month segments; contract upgrades may be impactful.
- In banking, inactivity and declining transaction behavior can trigger targeted retention (rewards, engagement campaigns).
- In SaaS, the model suggests engagement decline matters, but it also signals missing drivers; product telemetry and billing metadata could be necessary for stronger performance.

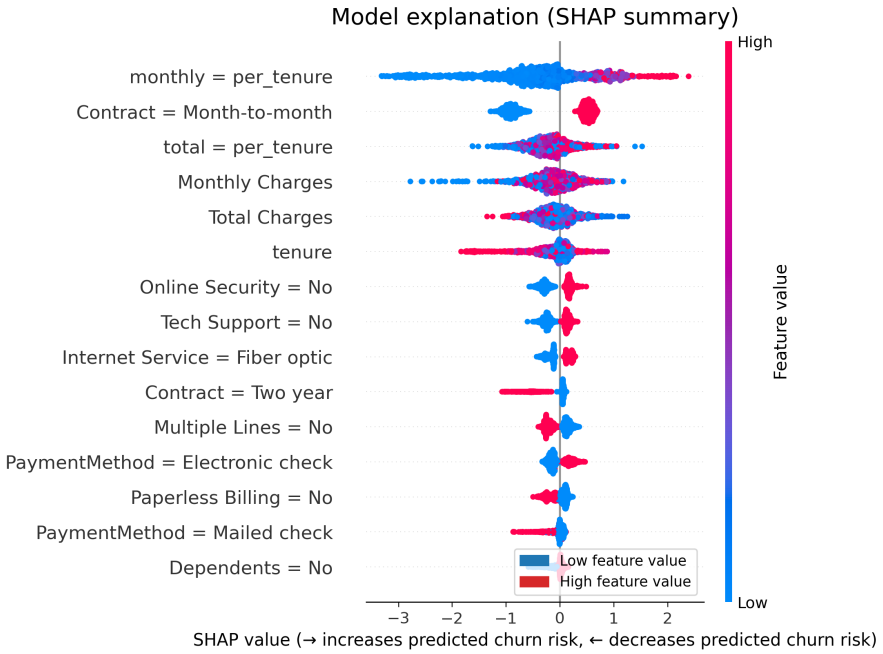

Figure 6: SHAP beeswarm plot for Telco. Tenure and contract-related features dominate churn risk.

8 Limitations

A principal limitation is the **SaaS dataset's inability to reach the target AUC threshold**. This is best understood as a data limitation rather than a modeling failure:

- **Small sample size:** with only 963 instances, variance is higher and category frequencies are less stable, limiting generalization.
- **Missing churn precursors:** the dataset does not include common pre-churn features such as days since last login, renewal failures, delinquency, refund requests, or explicit cancellation workflows.
- **Behavioral overlap:** engagement signals (viewing hours, downloads) can be similar for churned and retained users, producing distributional overlap and limiting separability.

As a result, even extensive hyperparameter tuning and stronger encodings cannot reliably push AUC beyond what the observed feature set supports. Future improvements would likely require richer temporal and billing-related features

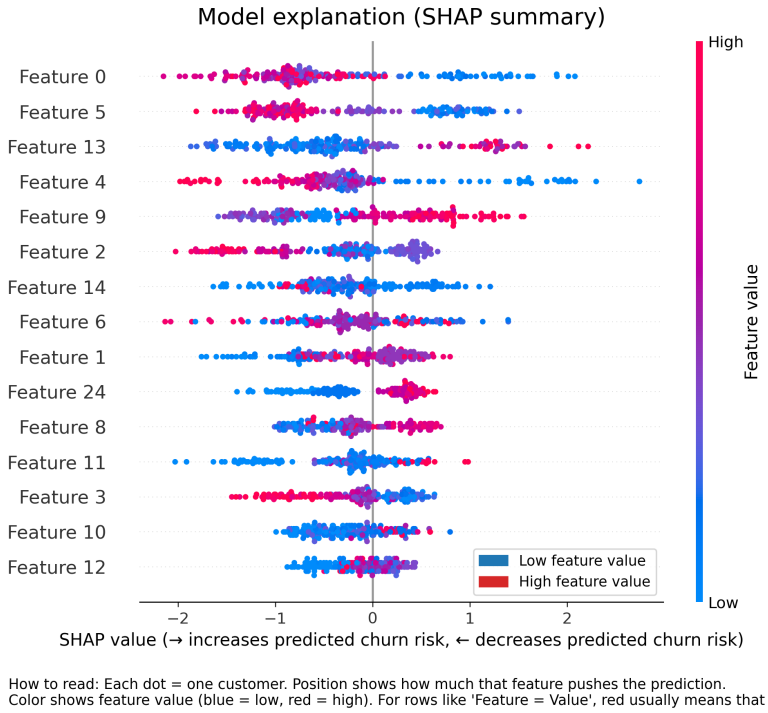

Figure 7: SHAP beeswarm plot for SaaS. Engagement features contribute, but effects are less separable than Telco/Banking.

(e.g., time-series modeling of engagement decay, renewal events, and payment friction), which are not available in the provided SaaS data.

9 Conclusion

We presented a unified churn prediction and explainability framework using XGBoost evaluated across three domains. Telco and BankChurners achieve AUC at or above the target threshold due to strong contract and transactional precursors, while the SaaS dataset exhibits an intrinsic performance ceiling due to limited sample size and missing pre-churn signals. SHAP explanations provide interpretable churn drivers that align with domain intuition and support practical retention strategies. Overall, this project demonstrates that the most effective path to robust churn prediction is a combination of strong models and high-quality domain instrumentation.

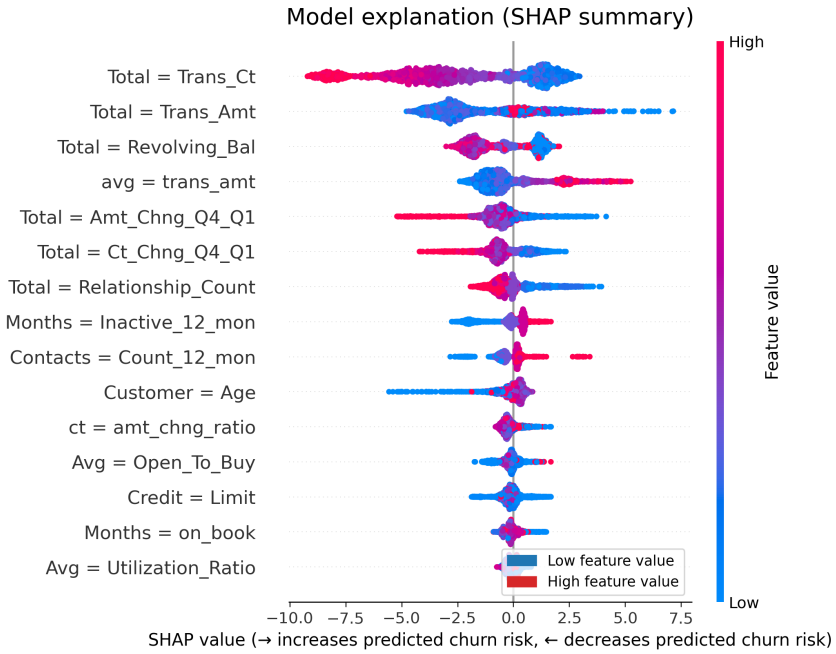

Figure 8: SHAP beeswarm plot for BankChurners. Inactivity and transaction behavior strongly drive attrition predictions.

Acknowledgements

This project was completed as part of coursework instructed by **Prof. Dr. Ruixiang Tang** at Rutgers University.

References

[1] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[2] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.