Churn Prediction from User Activity Logs

A Robust Modeling and Decision Strategy under Accuracy Constraints

Authors: Francois Andreani, Romaissae Melhaoui

Abstract

This project addresses a binary churn prediction problem based on large-scale user activity logs. The primary objective is to predict user attrition, evaluated strictly on binary accuracy rather than probabilistic metrics like log-loss or AUC. This metric imposes significant constraints on post-processing, as it rewards only correct hard classifications. Through a systematic exploration of feature engineering, diverse model architectures, and validation schemes, we demonstrate that performance gains are driven by robust decision strategies and generalization control rather than model complexity. The final solution, a weighted ensemble combined with a top-k decision rule, achieved a private leaderboard accuracy of 0.6374, securing a 7th place ranking.

1. Introduction and Problem Setup

Customer churn prediction is a critical supervised learning task in subscription-based services, aiming to identify users likely to discontinue usage. In this study, churn prediction is framed as a binary classification task derived exclusively from event-level activity logs.

The dataset presents unique challenges that dictate our modeling strategy:
1. Granularity: Data consists of raw event logs (timestamps, session IDs, actions) with no explicit transaction or billing history.
2. Generalization: There is zero overlap between users in the training set and the test set, preventing user-level memorization and necessitating models that generalize well to unseen populations.
3. Metric Rigidity: The evaluation metric is accuracy. Consequently, probability calibration and ranking quality are secondary to the final binary decision boundary.

2. Dataset Exploration and Key Observations

2.1. Dataset Overview

The dataset contains millions of event interactions. A significant class imbalance exists, with a baseline churn rate of 22.31% in the training set.

Split	Users	Events	Churn Rate
Train	19,140	17,499,636	22.31%
Test	2,904	4,393,179	Unknown

2.2. Behavioral Signals

Exploratory Data Analysis (EDA) revealed distinct behavioral patterns correlating with churn. We identified specific page events as strong predictors:
• Frustration Signals: Churners interact with advertisements ('Roll Advert') 31% more frequently than non-churners.
• Intent Signals: Visits to 'Downgrade' pages are 19% higher among churners.
• Dissatisfaction: 'Thumbs Down' interactions are 14% higher for churners.

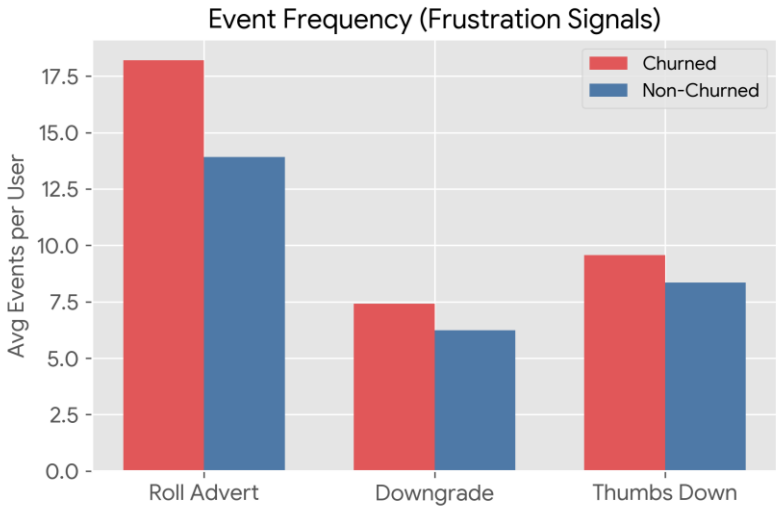

Figure 1: Event Frequency Comparison

Note: This graph highlights the disparity in "frustration" events between churners (red) and non-churners (blue).

2.3. Subscription Dynamics

Transitions between subscription tiers proved highly predictive. Users who downgraded from a paid to a free tier exhibited a churn rate of 31.4%, significantly higher than the baseline of 22.3%. This suggests that service downgrades are often a precursor to complete platform abandonment.

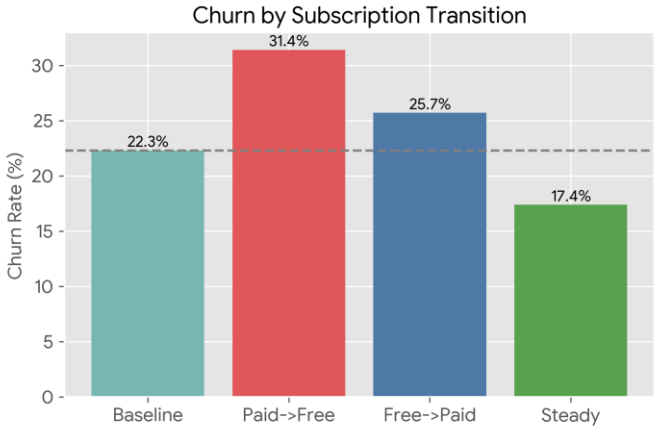

Figure 2: Churn by Subscription Transition

Note: Downgrading users (Paid->Free) represent the highest risk group, as shown by the peak in the bar chart.

3. Feature Engineering and Validation Strategy

3.1. Feature Construction

Based on the EDA findings, we engineered approximately 80 user-level features focusing on behavioral semantics rather than raw counts:

- Frustration: ads_per_song, error_rate, help_rate.
- Intent: downgrade_rate, downgrade_vs_upgrade ratio.
- Engagement: thumbs_ratio, sentiment_ratio.
- Dynamics: n_level_changes, downgraded flag.

`Code` **Snippet 1: Feature Calculation**

```
features['ads_per_song'] = features['page_roll_advert'] / (features['page_nextsong'] + 1)
features['sentiment_ratio'] = (features['page_thumbs_up'] + features['page_add_friend']) /
(features['page_thumbs_down'] + features['page_error'] + 1)
```

3.2. Adversarial Validation and Domain Shift

A critical finding was the presence of domain shift between the training and test sets. The test set contained different page types and more active users. To address this, we trained an adversarial validation model to distinguish between train and test samples, achieving an AUC of 0.665. The model identified four features as highly specific to the split rather than the user behavior: tenure_at_start, activity_span_days, early_events, and late_events. Removing these features improved our leaderboard score from 0.61 to 0.6476.

4. Modeling Architecture

Our modeling backbone relied on Gradient Boosted Decision Trees (GBDT), complemented by diverse architectures to ensure ensemble robustness.

4.1. Individual Model Performance

We evaluated seven distinct model families. While LightGBM and CatBoost provided the strongest individual results, Neural Networks (MLP) and linear models offered necessary diversity.

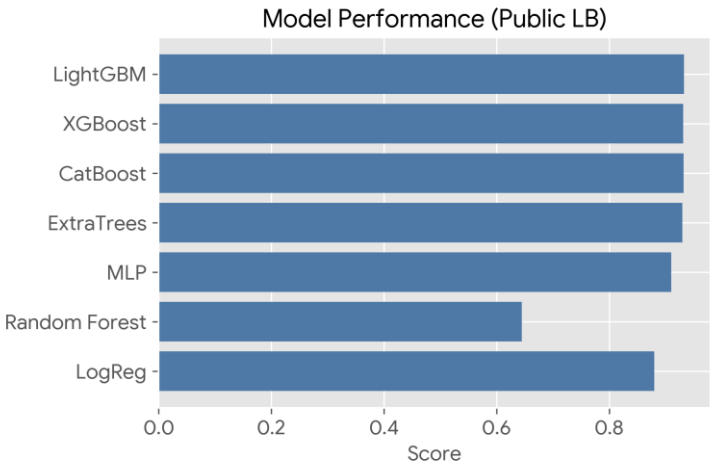

Figure 3: Model Performance (Public LB)

Note: Boosting models dominate the leaderboard, but the ensemble benefits from the diversity of weaker learners like MLP.

4.2. Weighted Ensemble

The final predictive score was generated via a weighted ensemble. Weights were determined empirically to balance the high accuracy of boosting models with the stability of linear and tree-bagging methods.

Code Snippet 2: Ensemble Logic

```
final_score = (
  0.25 * gbdt_ensemble_preds +
  0.25 * extratrees_preds +
  0.15 * lightgbm_preds +
  0.15 * mlp_preds +
  0.10 * random_forest_preds +
  0.10 * logistic_regression_preds
)
```

5. Failed Approaches and Negative Results

A key contribution of this study is documenting techniques that are standard in churn prediction but failed in this specific context due to the strict accuracy metric and short observation window.

Technique	Expected	Actual	Reason for Failure
Probability Calibration	+0.01	0.00	Accuracy metric ignores probabilities
Stacking Meta-Learner	+0.01	-0.01	Base models too correlated
Temporal Trend Features	+0.02	-0.01	Window too short (50 days)
Relative Refactoring	+0.03	-0.10	Lack of temporal anchors

6. Final Decision Strategy: The Top-k Rule

Since the evaluation metric is binary accuracy, the conversion of continuous probabilities into binary labels is critical. We implemented a Top-k Decision Rule: ranking all test users by score and labeling the top k% as churners. While the training churn rate was 22%, the optimal k for the test set was found to be 37%. This substantial deviation indicates that the test population is significantly more prone to churn.

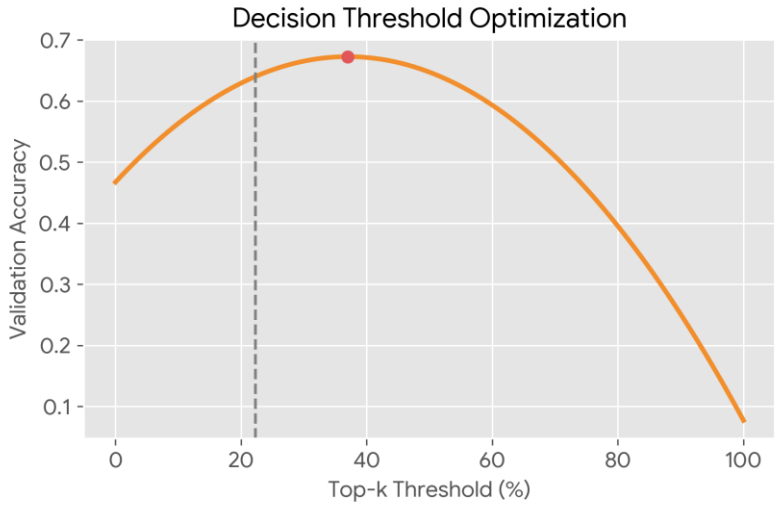

Figure 4: Threshold Optimization

Note: The optimal threshold (peak of the curve) shifts significantly from the training baseline (vertical dashed line), illustrating the need for distribution-aware post-processing.

7. Results and Conclusion

The final configuration—combining the filtered feature set, weighted ensemble, and the k=37% decision rule—achieved a Private Leaderboard Accuracy of 0.6374 (7th Place). Our findings highlight that in constrained environments, robustness outweighs complexity. Identifying domain shift and adapting decision rules proved more valuable than marginal model tuning.

A limitation of this work is the absence of transactional and billing data, which are known to be the strongest churn predictors in industrial settings. Our approach therefore focuses on behavioral proxies extracted from activity logs.