Marketing Multimodal banking(MDB) attrition project

Kshitij Nitin Patil (kp3113), Debbie Dai (fd2585), Saravanan Hari Baskaran (sh4621),
Xinyan Li (xl3469), Yuntong Ju (yj2564)

December 15, 2025

1. Introduction

1.1. Motivation

Customer attrition presents one of the most significant financial threats in the banking industry. According to a report from Capgemini, U.S. retail-bank retention fell from 78% in 2022 to 76% in 2023 (Patel, 2024). Even small declines in retention can translate into tens or hundreds of millions of dollars in lost revenue, as each retained customer contributes substantial lifetime value. Despite having access to extensive behavioral, geolocation, and customer-interaction data, many churn-risk models used in practice remain narrow in scope, typically rule-based or unimodal, and fail to capture how signals across customer channels interact over time. These limitations motivate the need for more comprehensive churn prediction frameworks that integrate heterogeneous data sources and account for temporal dynamics in customer behavior. We will use this model to design proactive retention strategies, such as targeted discounts and extra cash-back incentives, while prioritizing precision-recall optimization and feature importance to ensure interventions are both accurate and explainable.

1.2. Existing Work

Early customer-churn research in banking has largely relied on single-modal, structured-data models, including logistic regression, tree-based methods, and gradient boosting, which remain strong baselines on tabular churn datasets (*Zhang, 2024*). Time-to-event approaches using survival analysis, such as Random Survival Forests, extend this setting by predicting both whether and when churn occurs, outperforming classical Cox models in time-dependent accuracy (*Mustonen, 2025)*. However, these methods typically depend on aggregated transactional or CRM features and treat customer behavior as static snapshots. Related work on uplift modeling highlights similar limitations: while useful for campaign evaluation, uplift frameworks still rely on single-channel structured inputs and fail to capture the broader behavioral context of customers (*Belbahri et al., 2020*). More recent studies demonstrate the value of multimodal and sequential modeling, showing that customer behavior is shaped by signals across multiple sources and evolves. In particular, transaction-sequence modeling approaches treat financial activity as temporally ordered event sequences, and self-supervised sequence-representation methods such as CMLM-CoLES learn embeddings that capture both long-term behavioral patterns and short-term dynamics, consistently outperforming static feature-based approaches across downstream tasks (*Rudd et al., 2023; Yugay & Zaytsev, 2024*). Taken together, prior studies highlight two persistent limitations in the churn-prediction literature. First, most existing models continue to rely on single-modal inputs, despite evidence that customer behavior is shaped by rich behavioral, emotional, and spatial signals. Second, multimodal fusion and sequential modeling have each shown promise independently, but relatively few works integrate multimodal features with temporal patterns in a unified churn-prediction framework. As a result, current approaches often fail to fully capture the customer disengagement.

1.3. Problem Statement

Despite the availability of rich customer data, most churn-prediction systems in banking remain limited to single-modal or survival-based models that rely primarily on transactional history and treat customer behavior as static. As a result, temporal patterns and interaction signals that may

indicate early disengagement are often overlooked.

The problem addressed in this project is to develop a churn-prediction framework for retail banking that leverages multimodal behavioral data while **explicitly evaluating whether additional modalities provide incremental value beyond transactions alone**. This requires integrating heterogeneous and asynchronous data sources, capturing temporal dynamics, and addressing class imbalance. The objective is to enable early and high-confidence identification of at-risk customers to support targeted retention interventions, ensuring that added model complexity is justified by measurable gains in predictive and operational value.

1.4. Overview

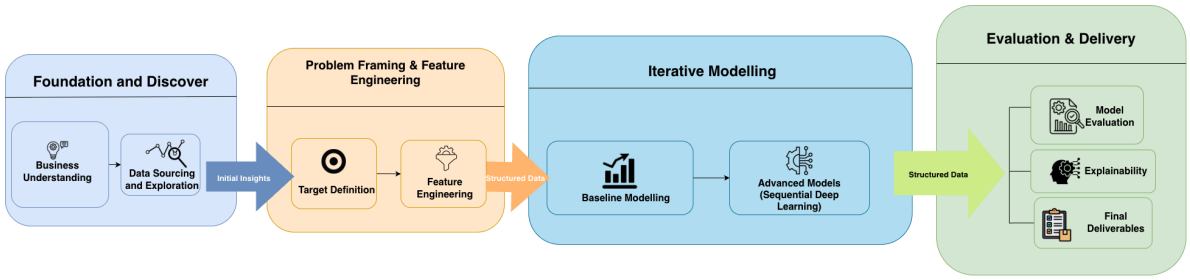

Figure 1: The Project Journey

The project begins with business problem formulation and exploratory analysis to clarify the operational definition of churn and to identify early warning signals embedded in customer behavior. We then construct a feature-engineering pipeline that integrates transactional activity, geospatial mobility, and customer dialogue interactions under strict temporal constraints to prevent data leakage. Building on these representations, we apply and compare strong tabular baselines with sequential learning approaches designed to capture temporal dynamics in customer engagement. Model performance is evaluated using both statistical metrics and business-oriented ranking measures, and the report concludes by examining model interpretability, practical deployment considerations, and implications for proactive retention strategies.

2. Methods

2.1. Data

2.1.1. Dataset Variant

In this project we use MBD-mini, a publicly available subset of the Multimodal Banking Dataset (MDB) introduced by Sber AI Lab. MBD-mini includes approximately 10% of all the clients in the full dataset by client IDwhile preserving its original temporal structure, multimodal composition, and behavioral characteristics. All Data is collected from the real world but partially anonymized through hashing to prevent privacy leakage.

2.1.2. Dataset Modalities

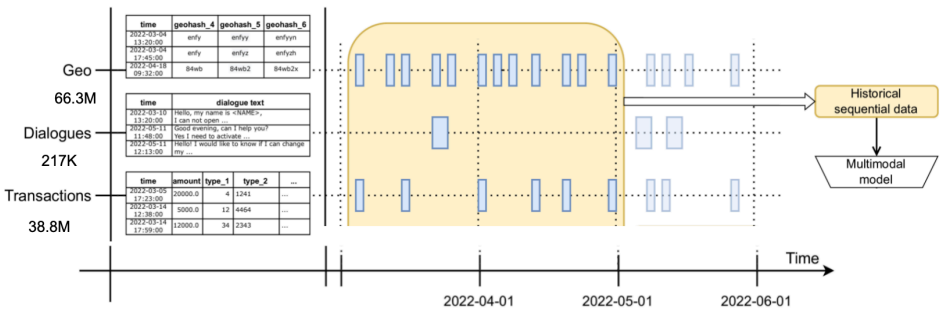

Figure 2: Data Structure Description

The dataset contains 3 primary data modalities, each providing complementary behavioral signals:

- **Transactional data (38.8M)**, capturing detailed financial activity including spending amounts, transaction types, and temporal patterns.
- **Geolocation data (66.3M)**, represented by time-stamped geohash events that describe customer mobility and location stability over time.
- **Dialogue data (217K)**, consisting of embeddings derived from customer–support interactions, reflecting engagement intensity.

MBD-mini spans 12 months of client activity and includes multimodal sequential signals together with monthly labels. These modalities are recorded as sequential event logs per client, allowing customer behavior to be analyzed as a temporally ordered process rather than static aggregates, consistent with recent advances in event-sequence modeling for financial behavior.

2.1.3. Dataset Structure and Scope

Records from all modalities are merged at the client level and aligned temporally. After preprocessing and aggregation, the resulting dataset contains approximately 236K client-month samples, spanning April 2022 to October 2022, and covers roughly 98,000 unique clients. Each client is represented by a sequence of monthly observations derived from their historical activity across all modalities. Transaction records with missing or undefined values are excluded during preprocessing to ensure data quality.

2.2. Exploratory Data Analysis

2.2.1. Transactions Modality

We conducted exploratory analysis of transactional behavior and attrition, including temporal patterns, churn stratification, and variance-based clustering of engagement trajectories. The reduced MBD dataset contains timestamped transactions with event subtypes, counterpart metadata, and monetary amounts, enabling both level- and trajectory-based analysis.

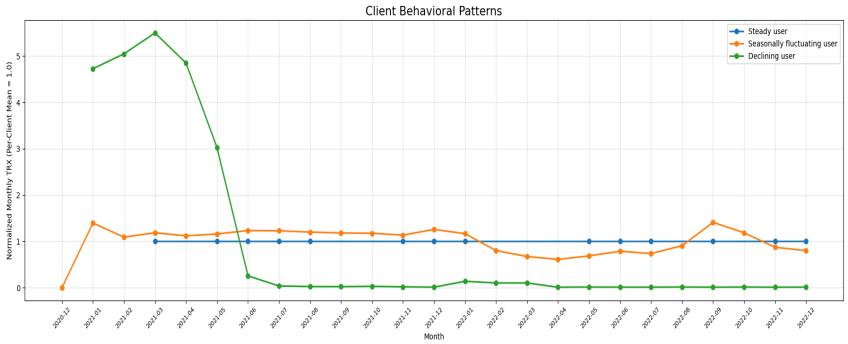

Figure 3: Client Behavior Patterns

A central finding from this EDA is that, as shown in Figure 3, **clustering by variance in normalized monthly transaction counts** reveals three dominant behaviors: **steady users, seasonally fluctuating users, and declining users**. These patterns indicate that attrition typically manifests as a gradual erosion of activity rather than an abrupt stop, suggesting that temporal dynamics carry a stronger signal than any single month's activity snapshot.

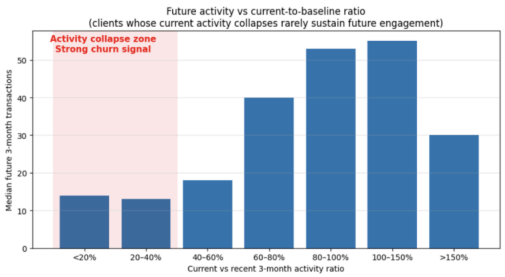

Figure 4: Future vs Current Activity Ratio

We further examined how short-term engagement decay relates to future behavior. Figure 4 shows a clear tipping-point effect: as transaction inactivity increases, churn rates rise sharply, and clients experiencing pronounced drops relative to their recent three-month baseline rarely recover prior engagement levels. This monotonic relationship reinforces the value of relative change features (e.g., activity ratios and declines) over raw transaction counts, which are highly skewed and provide limited separation.

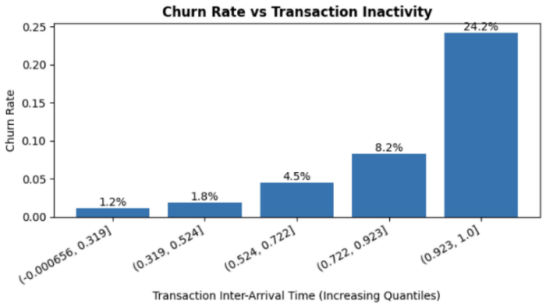

Figure 5: Churn Rate vs Transaction Inactivity

Together, these exploratory findings indicate that client attrition is a temporal process driven by progressively lengthening inactivity gaps and evolving engagement trajectories rather than static activity levels. In particular, relative behavioral changes, such as sustained inactivity or declining patterns (Figure 5), are far more predictive of churn than absolute transaction volumes, motivating the trajectory-based feature engineering and sequential modeling approaches described in later sections.

2.2.2. Geo Modality

The geo dataset contains timestamped location events with geohash identifiers and visit frequencies. Exploratory analysis revealed substantial temporal and coverage heterogeneity, including a ramp-up phase in early 2022 followed by stabilization, motivating restriction to periods with reliable geo availability.

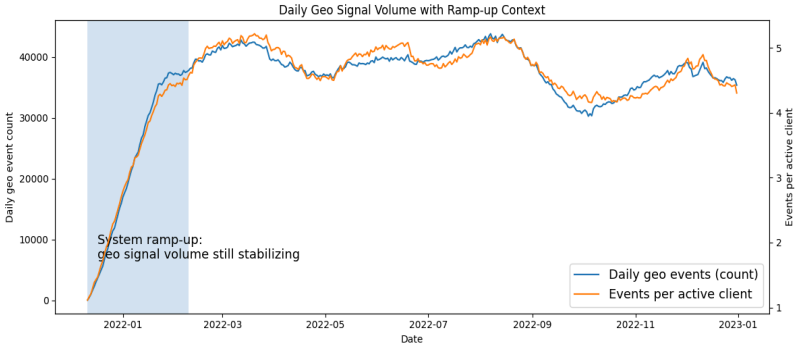

Figure 6: Geo Signal Volume with Ramp-Up

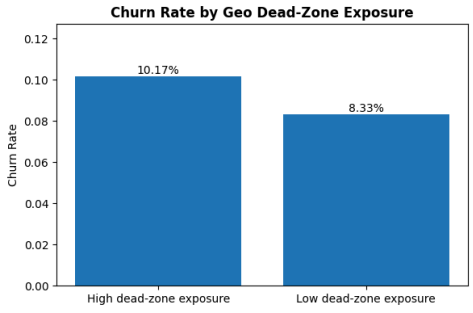

Figure 7: Churn Rate by Dead Zone

Beyond aggregate volume, we analyzed spatial engagement patterns such as mobility range (unique geohash exposure) and dead-zone time (proportion of visits in low-activity regions), both of which exhibit heavy-tailed distributions across clients. When evaluated against the true churn outcome, clients with high dead-zone exposure exhibit a higher churn rate than those with low exposure (Figure 7), linking degraded geo signals directly to disengagement risk. Together, these findings motivate the inclusion of geo-derived temporal features as complementary predictors beyond transactional behavior.

2.2.3. Dialog Modality

We analyzed timestamped client interaction logs aggregated at the monthly level to understand

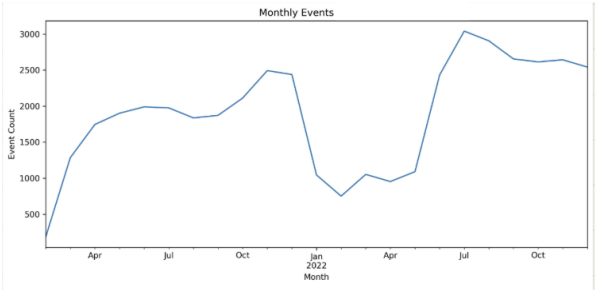

Figure 8: Monthly Event Volume Over Time

engagement dynamics and their relationship to downstream behavior. Figure 8 illustrates clear temporal non-stationarity in dialog activity: volumes rise early in the year, dip sharply between January and March 2022, and subsequently recover with sustained higher engagement.

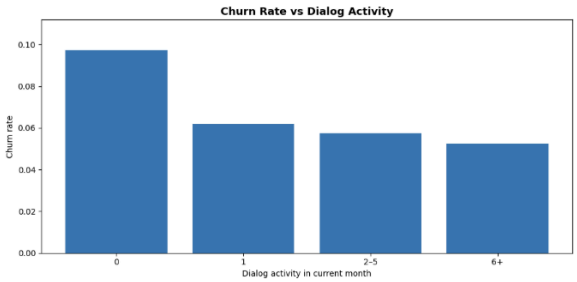

Figure 9: Churn Rate vs Dialog Activity

To assess behavioral relevance for attrition, we stratified clients by their dialog activity in the current month and compared observed churn rates, as defined in Section 2.3. As shown in Figure 9, churn rates vary systematically across dialog activity levels: clients with no dialog interactions exhibit the highest churn, while those with more frequent interactions show progressively lower churn.

This monotonic relationship suggests that dialog activity encodes meaningful engagement signals relevant to churn risk, motivating its inclusion as a complementary feature, such as interaction frequency, recency, and variability, in downstream modeling.

2.3. Analytics/Modeling Methods

2.3.1. Target Definition

We define customer churn as a severe and sustained decline in transaction activity relative to a customer's own historical baseline. For each reference month *M*, we compute the customer's baseline activity using the previous 6 months (*M−6* to *M−1*), then compare it to the customer's observed activity in the next 3 months (*M+1* to *M+3*). A customer is labeled as churned if their future activity falls below 20% of what would be expected from their baseline.
To be clearer, a customer-month (*c, M*) is labeled only if the customer has

1. At least 6 months of history before *M*,
2. At least 3 months of future observations after *M*
3. Baseline activity above a minimum low-activity threshold.

Client Eligibility criteria for modelling: Reference months are retained only if customers have a complete 6-month baseline window (M−6 to M−1) and a complete 3-month future window (M+1 to M+3). To ensure stable labeling, we require a minimum baseline activity of 4 transactions per month (averaged over the baseline period); customers below this threshold are excluded as no meaningful decline can be defined in these cases.

2.3.2. Churn Ratio and Label

We define the baseline average as the average transaction count over the past 6 months; expected activity is set to three times this baseline, and observed activity is the realized transaction count over the subsequent 3 months. We then compute a churn ratio as below: (with a small stabilizing constant to avoid degenerate cases),

$$\text{churn\_ratio}(c, M) = \frac{\text{actual\_future}(c, M)}{\text{expected\_future}(c, M) + \varepsilon}$$

A customer is labeled as churned if their churn ratio falls below a predefined threshold which we set to be 0.2. And observations with zero expected future activity (i.e., baseline average equal to zero) are excluded before labeling. Full calculation details are provided in Appendix F.

$$\text{churned}(c, M) = \begin{cases} \mathbf{1}, & \text{if churn\_ratio}(c, M) < \mathbf{0.20} \\ \mathbf{0}, & \text{otherwise} \end{cases}$$

2.3.3. Feature Engineering

Feature engineering for churn prediction follows strict temporal separation to prevent data leakage. All features are computed from a fixed 6-month historical window ending at the reference month *M*, while churn labels are derived exclusively from future observations. Full diagnostic procedures and validation results are reported in Appendix A, and Appendix E for detailed feature engineering explanation.

Category	Count	Description
Current Month	3	Activity snapshot at reference month
Recent Aggregates	8	Aggregated patterns over 6 months
Time-Series	22	Trend, momentum, and volatility
Ratio Features	1	Current vs historical comparison
Dialog Features	12	Customer service interactions
Geographic Features	16	Location-based patterns
Total	**62**	

Table 1: Engineered Features by Category

Feature Categories

We created 62 features and organized them into 6 categories based on their sources and temporal scope.

- **Current Month Features(3) -** Current month features capture the customer's activity state at the reference month, providing an immediate snapshot of engagement.
- **Recent Aggregate Features(8) -** These features aggregate customer behavior across the entire 6-month feature window, capturing longer-term patterns.
- **Ratio Features(1)** - The primary trend indicator compares current activity to the 6-month average. It directly measures the direction and magnitude of activity change relative to established patterns.
- **Time Series Features(22) -** Time-series features capture temporal dynamics, including trends, momentum, and volatility across the 6-month window.

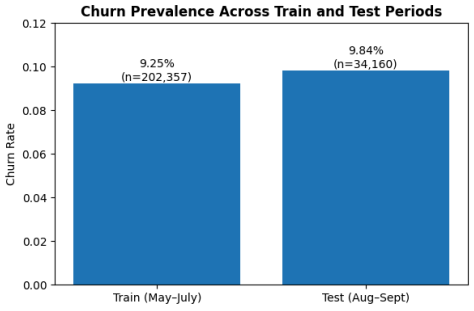

Figure 10: Churn Across Train and Test Periods

- **Dialog Features(12) -** Customer service interactions are captured through dialog features computed over the 6-month window.
- **Geographic Features(16) -** Location-based features are derived from geo-tagged transactions to reveal diversity and concentration relative to location information.

At this step, before entering the model, we split the data for train and test. Full details can be found in *Appendix G.* As shown in Figure X, churn prevalence is comparable across splits, indicating a stable label distribution over time.

Data Coverage: Transaction data is complete for all customers (88,076). Dialog and geographic modalities exhibit expected sparsity, with dialog data missing for ~52% of customers and geo data missing for ~27%. As a result, only 37% of customers have full multimodal coverage, while 16% have transaction data only. Churn rates are higher among customers with missing modalities: 11.6% vs. 6.9% for those without vs. with dialog data, and 11.9% vs. 8.4% for those without vs. with geo data.

Key takeaway: Missing dialog and geographic data is structurally common and strongly associated with higher churn risk.

2.3.4. Static Baseline Models

Building on the aggregated multimodal features described above, we first evaluate static baseline models that operate on fixed-length monthly snapshots. This reflects common practice in churn prediction, where risk is assessed from summarized historical behavior rather than raw event sequences. Monthly aggregates capture stable engagement signals, such as recency, frequency, intensity, and cross-channel usage, while reducing noise from irregular event timing. We include both linear and nonlinear baselines: logistic regression serves as a transparent linear reference, while gradient-boosted tree models (XGBoost, LightGBM) capture higher-order interactions across transactional, geospatial, and dialogue features. All baseline models are trained using **a time-based split by reference month with client-disjoint partitions, ensuring training months precede test months** to prevent information leakage across users, and **class-weighted loss functions** are applied to mitigate severe class imbalance.

Overall, these baselines validate that the engineered multimodal features carry meaningful predictive signals and provide a competitive non-sequential reference for evaluating the added value of temporal modeling. Implementation details appear in *Appendix B*.

2.3.5. Sequential Modelling

While static baselines capture level-based engagement signals, they cannot represent behavioral trajectories such as gradual disengagement, recovery, or sustained volatility that often precede attrition. Because churn is inherently temporal, predictive signal frequently lies in how customer behavior evolves rather than in any single month's snapshot.

To capture these dynamics, we adopt a sequential formulation in which each customer is represented as an ordered sequence of monthly behavioral summaries. This allows the model to learn temporal dependencies using the same feature definitions as the static baselines, isolating the value added by explicit temporal modeling. And here we use the Temporal Convolutional Network (TCN), which employs causal, dilated convolutions to model multi-month behavioral patterns while preventing the use of future information. Sequential models are trained under the same leakage-controlled, time-aware protocol as static models to ensure fair comparison. Full architectural and training details, including hyperparameters and alternative sequential baselines (e.g., GRU), are provided in *Appendix B*.

3. Preliminary Results

3.1. Evaluation Protocol

All models are evaluated under a leakage-controlled protocol designed to reflect operational churn prediction. For each reference month t, churn risk scores are generated using only customer behavior observed up to month t, ensuring temporal truncation of features.

Importantly, training and evaluation are performed on **disjoint sets of customers**, such that no client appears in both the training and test splits. This client-level separation prevents information leakage across time while allowing evaluation across multiple reference months. Model outputs are treated as continuous risk scores and evaluated primarily in a ranking-based framework, reflecting the business objective of prioritizing a limited subset of customers for proactive retention. This protocol is applied consistently across all models to ensure fair comparison.

3.2. Evaluation Metrics

Model performance is evaluated using both threshold-free and threshold-based metrics to reflect ranking quality and operational relevance under class imbalance. **ROC-AUC** measures overall discrimination, while **PR-AUC** focuses on performance on the minority churn class. **Precision**, **Recall**, and **F1-score** are reported at calibrated decision thresholds to capture the trade-off between churn identification and false alerts.

To directly assess targeting effectiveness, we additionally evaluate **cumulative gain** and **lift** curves. Let N denote the total number of customers in the evaluation set, and let π(i) index customers sorted in descending order of predicted churn risk. For a given cutoff K, corresponding to targeting the top K% of customers by predicted risk, cumulative gain and lift is defined as:

$$Gain(K) = \frac{\sum_{i=1}^{KN} 1\{y_{\pi(i)} = 1\}}{\sum_{i=1}^{N} 1\{y_i = 1\}}, \quad Lift(K) = \frac{Gain(K)}{K}$$

These metrics align evaluation with realistic retention budgets, where only a fixed fraction of customers can be targeted.

3.3. Comparison of Static and Sequential Models

Metric	Logistic Regression	XGBoost	LightGBM	TCN(Seq2Seq)	GRU
Precision	0.5302	0.5442	0.5446	**0.5675**	0.5205
Recall	0.6108	0.6203	**0.6313**	0.6168	0.6311
F1-Score	0.5452	0.5625	0.5795	**0.5911**	0.5648
PR-AUC	0.5120	0.5252	0.5257	**0.5460**	0.5502
ROC-AUC	0.9120	0.9121	0.9130	0.9130	0.905

Table 2. Static Models vs TCN

Table 2 compares Logistic Regression, XGBoost, LightGBM, a Temporal Convolutional Network (TCN), and a GRU. Logistic regression establishes a strong linear baseline (ROC-AUC = 0.912), confirming that the aggregated multimodal features capture substantial predictive signals. However, all nonlinear models, tree-based and sequential, consistently outperform the linear baseline on PR-AUC and F1-score, indicating that modeling nonlinear interactions is beneficial for churn prediction. Among static models, LightGBM is the strongest, achieving a recall of 0.631 and F1-score of 0.580 after threshold calibration. Sequential models (TCN and GRU) deliver modest additional gains in precision and PR-AUC, with TCN achieving the highest F1-score (0.591) and GRU the highest PR-AUC (0.550). Despite these improvements, ROC-AUC remains effectively unchanged across LightGBM and TCN, suggesting that while temporal modeling refines ranking quality in the imbalanced regime, its incremental benefit over well-tuned nonlinear static models is limited in this setting.

These results indicate that **well-designed static models trained on carefully engineered aggregated features are sufficient for churn prediction in this setting**. The marginal improvements offered by sequential modeling do not justify the additional modeling complexity, training cost, and deployment overhead. Consequently, LightGBM is selected as the preferred modeling approach for this task.

Metric	Trx	Trx+Geo	Trx+Dialog	Trx+Dialog+Geo
Precision	0.5120	0.5231	0.5141	0.5446
Recall	0.6208	0.6307	0.6226	**0.6313**
F1-Score	0.5620	0.5719	0.5632	0.5795
PR-AUC	0.5123	0.5254	0.5250	0.5257
ROC-AUC	0.8901	0.9100	0.9065	0.9130

Table 3. Performance of LightGBM with Different Feature Modalities

Table 3 reports LightGBM performance as additional feature modalities are incorporated. Using transaction features alone yields strong baseline performance (ROC-AUC = 0.890, PR-AUC = 0.512), indicating that transactional behavior captures most churn-related signals. Adding geographic features improves global discrimination (ROC-AUC = 0.910, PR-AUC = 0.525) and modestly increases F1-score, suggesting that coarse spatial context provides complementary information.

Dialog features primarily improve sensitivity rather than selectivity: the Trx+Dialog configuration slightly increases recall (0.623 vs. 0.621) but offers limited gains in precision, resulting in comparable F1-score and PR-AUC. Combining dialog and geographic features yields the strongest overall performance, achieving the highest recall (0.631), F1-score (0.580), and ROC-AUC (0.913), while also improving precision relative to transaction-only features.

Overall, modality augmentation does not uniformly improve performance. **Geographic features consistently enhance global ranking metrics and provide clear additive value**, while dialog features primarily increase recall at the expense of precision. As a result, incorporating dialog information is beneficial only in settings where maximizing coverage of at-risk customers is prioritized, whereas geographic features should be included as a default enhancement to transaction-based models.

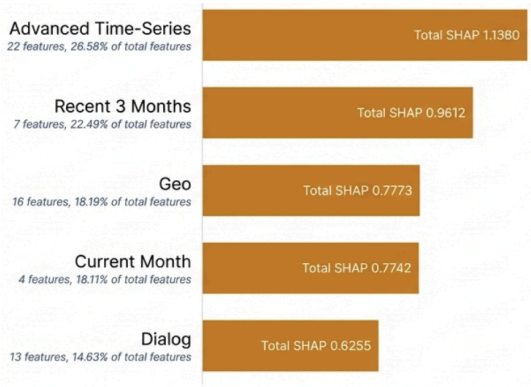

Figure 11: SHAP of Features

Feature group attribution analysis indicates that transaction-derived features, particularly advanced time-series and recent aggregate features, dominate model predictions, while geographic features provide meaningful secondary context, and dialog features contribute more modestly. This aligns with earlier modality ablation results and helps explain why static models achieve strong performance. A more detailed instance-level interpretability analysis is provided in Appendix C.

3.4. Ranking-Based Evaluation and Targeting Effectiveness

To assess practical impact, we compare the strongest static baseline (LightGBM) against the TCN using cumulative gain and lift curves. These curves focus on the operational setting in which only a limited fraction of customers can be targeted for retention. As shown in Figure 9, targeting the top 20% of customers ranked by LightGBM captures approximately 80% of observed churners, closely matching the performance of the TCN. The two curves overlap across most of the ranking range, indicating that both models prioritize high-risk customers with comparable effectiveness. Similarly, the lift

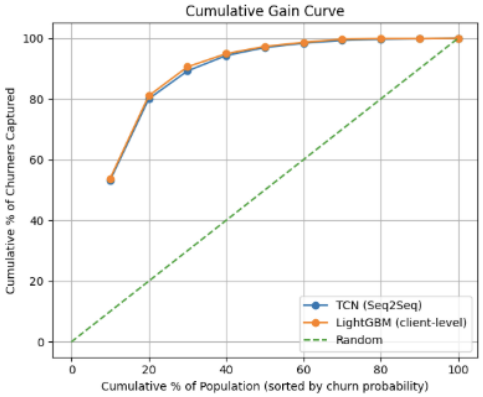

Figure 12: Gain Curve

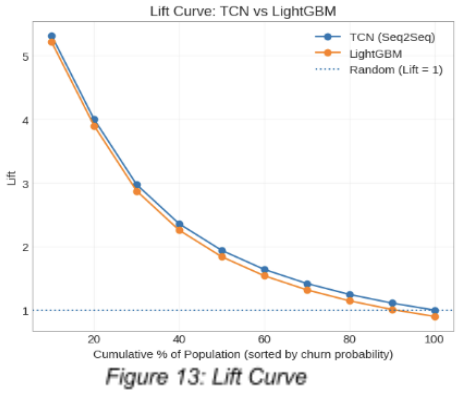

Figure 13: Lift Curve

curves in Figure 10 show that targeting the top 10% of customers yields nearly a five-fold increase in churn capture relative to a random baseline for both approaches. Across all targeting thresholds, LightGBM achieves a lift that is nearly indistinguishable from the TCN, demonstrating that the static model concentrates churn risk just as effectively in the highest-risk segments. Overall, the ranking-based analysis reinforces earlier findings: **while sequential modeling offers marginal refinements, a well-tuned static LightGBM model delivers equivalent operational value**.

4. Discussion

4.1. Conclusion

This work evaluates a multimodal churn prediction framework that integrates transactional behavior with geographic and dialog signals, emphasizing both predictive performance and operational relevance. Across all experiments, transaction-centric features, augmented with temporal aggregates and time-series descriptors, emerge as the dominant source of predictive signal. Using these representations, the strongest static baseline, LightGBM, achieves robust performance with a ROC-AUC of approximately 0.91 and a PR-AUC of 0.53, indicating strong ranking capability in an imbalanced churn setting.

Modality analysis shows that **geographic features provide consistent additive value**, improving global discrimination and ranking quality, while **dialog features contribute positively but more modestly**, primarily by increasing recall and expanding coverage of at-risk customers at the cost of reduced precision. Feature group attribution further supports this finding, with advanced time-series and recent aggregate features dominating model explanations, geographic features providing meaningful secondary context, and dialog features acting as a complementary but weaker signal.

Overall, these results indicate that well-engineered static models capture most of the actionable churn signal, with non-linear tree-based approaches outperforming linear baselines and achieving performance comparable to sequential models. Ranking-based evaluations further show that both LightGBM and TCN are effective at prioritizing high-risk customers, capturing roughly **80% of churners within the top 20% of ranked clients**. Modality-specific analysis suggests that **geographic and dialog features provide incremental but uneven benefits**, improving ranking quality or recall depending on the operating point, rather than uniformly dominating transaction-only models. Together, these findings highlight the importance of evaluating model class, feature modalities, and operational objectives jointly, rather than assuming increased complexity or multimodality is always beneficial.

4.2. Future Work

Future work could enhance operational robustness by evaluating performance on more recent periods (e.g., 2024–2025) to assess generalization under data shift and determine the need for recalibration. While static models capture most actionable signals in this setting, alternative temporal formulations may be explored in scenarios with longer histories or different churn definitions, including attention-based or survival-oriented approaches. Further interpretability could focus on identifying which time periods or behavioral transitions drive risk predictions. Finally, extending the framework from prediction to prescription via uplift or causal modeling would enable more targeted, ROI-driven retention strategies.

5. References

Belbahri, M., Murua, A., Gandouet, O., & Partovi Nia, V. (2020). *Qini-based uplift regression*. arXiv. https://arxiv.org/abs/1911.12474

Mollayev, D., Kostin, A., Postnova, M., Karpukhin, I., Kireev, I., Gusev, G., & Savchenko, A. (2024). *Multimodal banking dataset: Understanding client needs through event sequences*. arXiv. https://arxiv.org/abs/2409.17587

Mustonen, V. (2025). *Customer churn estimation in the banking sector with random survival forest* (Master's thesis). Tampere University.

Patel, V. (2024, September 25). *Creating a loyal experience for customers in banking*. Capgemini. Retrieved October 24, 2025, from https://www.capgemini.com/us-en/insights/expert-perspectives/creating-a-loyal-experience-for-customers-in-banking/

Rudd, D. H., Huo, H., Islam, M. R., & Xu, G. (2023). *Churn prediction via multimodal fusion learning: Integrating customer financial literacy, voice, and behavioral data*. arXiv. https://arxiv.org/abs/2312.01301

Yugay, A., & Zaytsev, A. (2024). *Uniting contrastive and generative learning for event sequences models*. arXiv. https://arxiv.org/abs/2408.09995

Zhang, K. (2024). *Predicting customer churn in banking industry using machine learning* (Master's thesis). Tilburg University.
Link to the codebase :- https://github.com/DebbieDai07/MDB_Attrition

Appendix A: Feature Diagnostics and Validation

To ensure that all engineered behavioral features were reliable, leakage-free, and suitable for downstream modeling, we conducted a six-point diagnostic framework combining temporal leakage tests, distributional analyses, redundancy checks, and stability evaluations. This appendix summarizes key validation steps and presents supporting plots.

1. Leakage Prevention and Temporal Integrity

a. Global "Recency Blackout" Test

We evaluated whether recent-month information inadvertently leaked future signals by retraining the LightGBM baseline with the *entire final month of features removed*. The ROC-AUC decreased only marginally (0.912 → 0.895), indicating that predictive power stems from stable behavioral patterns rather than short-horizon leakage. Figure A1 reports this comparison.

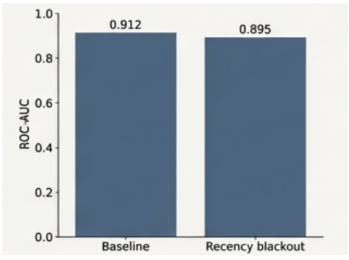

Figure A1: Recency Blackout

b. Per-Feature Lag-Swap Audit

To further confirm temporal integrity, each feature f_t was replaced with its lagged version f_{t-1} and the change in AUC was measured. If a feature contained leakage, replacing it with its prior-month version would cause large performance drops. Across all 62 features, AUC degradation was negligible (largest $\Delta AUC \approx 0.001$), demonstrating that no individual feature encodes future information as seen in *Fig A2*.

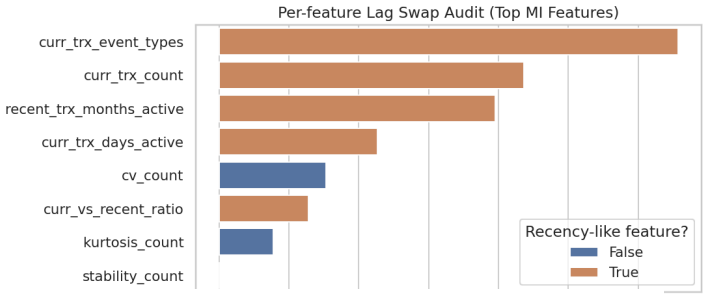

Figure A2: Per-Feature Lag-Swap Audit

2. Feature Quality, Stability, and Predictive Health Checks

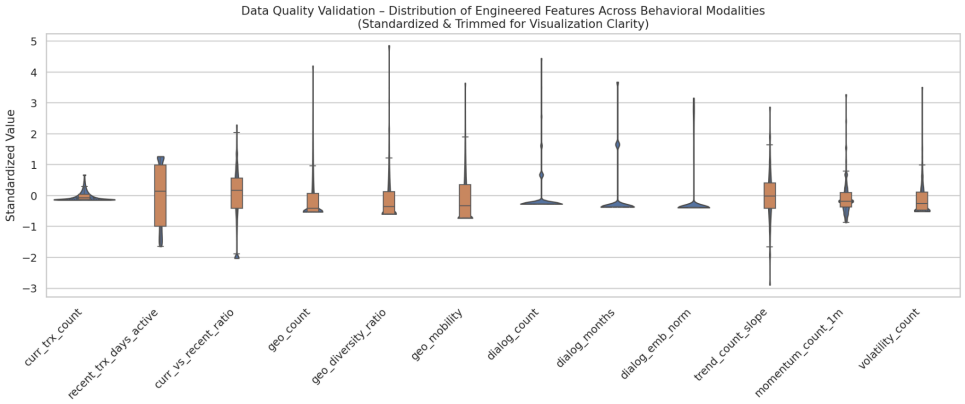

Figure A3: Distribution of Engineered Features

a. Distributional Consistency
All engineered features exhibited well-behaved empirical distributions with systematically handled outliers and missing values. This ensures that downstream model behavior is not driven by artifacts. *Fig A3* provides boxplots illustrating distributional regularity across churn and non-churn cohorts.

b. Predictive Signal Verification
Across most behavioral groups, churners and non-churners show meaningful distributional separation, confirming that features contribute real predictive signal rather than noise. This aligns with high permutation importance and SHAP attribution scores.

c. Temporal Stability of Behavioral Patterns
Trend-level diagnostics confirm that feature trajectories remain stable across the 18-month observation window. Stability indicates that learned patterns reflect consistent customer behavior rather than idiosyncratic fluctuations or data-collection artifacts.

d. Redundancy and Multicollinearity Checks
Variance Inflation Factor (VIF) and mutual information (MI) analysis indicate low redundancy across features. This reduces instability and prevents inflated model variance, particularly important when training deep sequential models.

Appendix B: Model Implementation Details

This appendix summarizes the key implementation choices for all models evaluated in the study. The goal is to provide sufficient detail for reproducibility while keeping the description model-agnostic and focused on principled design decisions rather than exhaustive parameter listings.

1. **Data Representation and Temporal Splits**

   All models operate on the same monthly aggregated multimodal features described in the Feature Engineering Section. For each client c and month t, we construct a feature vector

   $$x_t^{(c)} \in R^F,$$

   and a binary churn label $y_t^{(c)} \in \{0, 1\}$ indicating whether the client churns within a fixed 3-month window after month t.

   Static baselines consume a single snapshot per client (the reference month), while sequential models consume the entire history:

   $$X^{(c)} = (x_1^{(c)}, \ldots . x_{T_c}^{(c)}).$$

   Train/validation/test splits are time-aware and client-disjoint: training months strictly precede validation and test months, and each client appears in exactly one split. Feature standardization (mean and variance) is computed on the training split and reused for validation and test.

   Dataset summary:

   - **Train:** 202,357 samples (**9.25% churn**)
   - **Test:** 34,160 samples (**9.84% churn**)
   - **Number of features:** 62
     Label horizon: next 3 months
   - **Observation granularity:** monthly

   All models consume the same 62 aggregated multimodal features, derived from transactional, geolocation, dialogue, and engineered temporal signals. No explicit feature selection is applied in the main modeling pipeline. Instead, model capacity is controlled through regularization (logistic regression), tree constraints (LightGBM), and architectural inductive bias (TCN).

2. **Static Baseline Models**

   All static models approximate the conditional churn probability given a reference-month snapshot,

   $$\widehat{p}(y = 1 \mid z) = f_\theta(z),\ z \in R^F,$$

   where z is the aggregated feature vector for that month and f_θ denotes a parametric predictor.

   - **Class imbalance handling:**
     The positive (churn) class is reweighted using a factor proportional to the inverse class frequency. In the final training split, this corresponds to a **positive class weight of 9.81**.
   - **Hyperparameter tuning:**
     Hyperparameters are selected via **time-aware validation**, optimizing PR-AUC. For LightGBM, early stopping on validation PR-AUC is used to determine the optimal number of boosting iterations.
   - **Decision threshold:**
     A fixed operating threshold of **0.337** is selected on the validation set and applied consistently across models to enable fair comparison of precision–recall trade-offs.

   a. Logistic Regression:

   We train an ℓ_2-regularized logistic regression model

$$\widehat{p}(y = 1 \mid z) = \sigma(w^T z + b),$$

with loss

$$L_{LR} = \sum_i w_y \, BCE(y_i, \widehat{p}_i),$$

where $w_1 \propto N_0/N_1$ re-weights the minority churn class and $w_0=1$. Features are standardized using a StandardScaler and the model is optimized with a quasi-Newton solver until convergence. Regularization strength is selected via time-aware validation. Other than those, no specific tuning is applied.

b. Gradient-Boosted Trees (XGBoost, LightGBM).

Both XGBoost and LightGBM model $\widehat{p}(y = 1 \mid z)$ as an additive ensemble of decision trees,

$$f_\theta(z) = \sum_{m=1}^{M} \eta_m h_m(z),$$

trained with a logistic objective and tree-specific regularization (depth / number of leaves, shrinkage, subsampling). Class imbalance is handled via a positive-class weight scale_pos_weight $\approx N_0/N_1$. Hyperparameters such as tree depth, number of boosting iterations, learning rate, and subsampling ratios are tuned using rolling, time-aware validation; early stopping on validation PR-AUC was applied to prevent overfitting.

Technical Details

Model Type	LightGBM Gradient Boosting Decision Trees
Training Algorithm	Gradient Boosting Decision Tree (GBDT)
Number of Features	62
Number of Trees	8 (early stopping at iteration 8)
Learning Rate	0.05
Max Leaves	31
Feature Fraction	0.8
Bagging Fraction	0.8
Class Weight	9.81 (to handle class imbalance)
SHAP Sample Size	1,000 test samples
SHAP Method	TreeExplainer (exact SHAP values for tree models)

Figure A4: Tuning on LightGBM

3. Sequential Architectures

Sequential models operate on the full monthly behavioral history of each client. For client c, let the observed sequence be

$$X^c = (x_1^{(c)}, \ldots x_{l_c}^{(c)}), \qquad x_t^{(c)} \in R^F,$$

where $t \in \{1, \ldots, \ell_c\}$ indexes calendar months (time steps) in the client's history, F is the number of aggregated multimodal features per month, and ℓ_c is the true sequence length for client c.

The churn probability is modeled as

$$p(y = 1 \mid X^{(c)}) = f_\theta(x_1^{(c)}, \ldots x_{l_c}^{(c)}).$$

For minibatch training, sequences are padded to a common maximum length T_{max}. A binary mask

$$m_t^{(c)} = 1(t \leq l_c)$$

ensures that the loss is computed only on valid (unpadded) timesteps:

$$L(\theta) = \sum_c \sum_{t=1}^{T_{max}} m_t^{(c)} f_\theta(x_1^{(c)}, \ldots x_{l_c}^{(c)})$$

where $\ell(\cdot, \cdot)$ is the binary classification loss.

At the dataset level, customer histories are represented as a padded tensor:

$$X \in R^{N \times T \times F},$$

where N is the number of customers, T is the length of the observed behavioral history, and F is the number of aggregated features per month. This representation preserves temporal ordering while ensuring a controlled comparison with static baselines: both model families consume identical monthly features, but only sequential models can exploit **dependencies across timesteps**.

a. Gated Recurrent Units:

Before introducing the Temporal Convolutional Network (TCN), we also experimented with a Gated Recurrent Unit (GRU) architecture, as recurrent models are a natural choice for monthly longitudinal data. GRUs are widely used in sequential prediction tasks because they:

(1) maintain a hidden state that summarizes past behavior,
(2) adaptively gate information to capture long-term dependencies, and
(3) handle variable-length sequences without requiring fixed temporal windows.

The GRU uses a bidirectional gated recurrent unit to encode the sequence:

$$h_t = GRU(x_t, h_{t-1}), \quad \tilde{h_t} = [\overrightarrow{h_t}; \overleftarrow{h_t}].$$

A time-distributed linear head produces logits at each timestep,

$$s_t = w^\top \tilde{h_t} + b,$$

so that $\tilde{p_t} = \sigma(s_t)$ approximates the probability of churn for the horizon anchored at month t.

Training uses per-timestep supervision with a class-weighted binary cross-entropy:

$$L_{GRU} = \sum_c \sum_{t=1}^{T_c} w_{y_t^{(c)}} BCE(y_t^{(c)}, \widehat{p_t^{(c)}}),$$

where w_1 is capped to avoid unstable gradients under extreme imbalance. For evaluation, we report client-level scores using the final timestep of each sequence, aligning with the operational task of scoring the most recent month.

b. Temporal Convolutional Network (TCN)

Given the shared sequence representation described above, we now detail the TCN architecture used to model long-range behavioral dependencies. The model employs causal, dilated convolutions with residual blocks. For layer ℓ, with dilation d_ℓ and kernel size k, the hidden state is:

$$\mathbf{h}_t^{(\ell)} = \mathbf{h}_t^{(\ell-1)} + g\left(\sum_{j=0}^{k-1} \mathbf{W}_j^{(\ell)}\, \mathbf{h}_{t-jd_\ell}^{(\ell-1)} + \mathbf{b}^{(\ell)}\right),$$

where g(·) denotes a nonlinear transformation (ReLU) followed by normalization and dropout. Out-of-range indices are zero-padded to preserve causality. Dilation grows exponentially,

$$d_\ell = 2^{\ell-1},$$

allowing a large temporal receptive field with a small number of layers.

After the final TCN block, the network outputs a contextualized sequence

$$\mathbf{H} = \{\mathbf{h}_1^{(L)}, \ldots, \mathbf{h}_T^{(L)}\}.$$

To ensure temporal validity, we extract the representation at the final observed timestep for each client:

$$\mathbf{h}_c^{final} = \mathbf{h}_{\ell_c}^{(L)},$$

where ℓ_c is the last month with available history for client c.
A linear prediction head maps this representation to a churn score:

$$s^{(c)} = \mathbf{w}^\top \mathbf{h}_c^{final} + b, \qquad \hat{y}^{(c)} = \sigma(s^{(c)}).$$

To address the severe class imbalance in churn prediction, the TCN is trained using focal loss:

$$\mathcal{L}_{TCN} = -\sum_c \left[\alpha(1-\hat{y}^{(c)})^\gamma y^{(c)} \log \hat{y}^{(c)} + (1-\alpha)(\hat{y}^{(c)})^\gamma (1-y^{(c)}) \log(1-\hat{y}^{(c)})\right],$$

where α controls class weighting and γ > 0 down-weights easy examples. This objective emphasizes hard positive cases, which is critical in settings where only a small fraction of customers churn

Final TCN configuration:

- Hidden dimension: 128
- Number of layers: 2
- Kernel size: 2
- Dropout: 0.3
- Optimizer: AdamW
- Learning rate: 0.0005
- Weight decay: 0.001
- Loss function: Focal loss (α = 0.25, γ = 2.0)

Appendix C: Interpretability

Predictive accuracy alone is insufficient for actionable churn management; stakeholders must understand **which behavioral patterns the model relies on** and **why a specific client is flagged as high risk**. To address this, we analyze the selected **LightGBM model** through two complementary interpretability lenses:

(1) **Global feature importance**, which reveals population-level drivers of churn risk, and
(2) **Local, client-specific attributions**, which explain individual predictions.

Both analyses are based on **TreeSHAP**, which provides an exact additive decomposition of LightGBM predictions into feature-level contributions. This enables direction-aware, instance-level explanations while remaining consistent with the model's nonlinear structure, allowing churn risk to be interpreted transparently without reliance on complex sequential architectures.

Global Attribution: What Behavioral Patterns Drive Churn Across the Population?

To understand which behavioral signals drive churn risk at the population level, we analyze the selected **LightGBM model** using two complementary global interpretability approaches. Together, these methods quantify both the magnitude and direction of feature contributions while remaining consistent with the model's decision structure.

(1) **Permutation-based importance** measures a feature's marginal contribution to predictive performance by quantifying the degradation in ranking quality when the feature is randomly permuted across clients. This provides a global, model-agnostic estimate of feature relevance but does not capture directionality or instance-level effects

(2) **TreeSHAP** decomposes each LightGBM prediction into additive feature attributions, enabling direction-aware explanations that are locally accurate and globally consistent. Aggregating TreeSHAP values across clients yields a population-level view of which features most strongly influence churn predictions.

Permutation Importance:

For each feature f, permutation importance is defined as the reduction in ranking performance after permuting the feature across all clients:

$$Importance(f) = AUC_{baseline} - AUC_{permute(f)},$$

Larger values indicate that disrupting the feature substantially degrades predictive quality, implying that the feature encodes a meaningful behavioral signal. While this metric captures global dependence, it does not reveal how or in which direction a feature influences individual predictions.

SHAP for LightGBM

Permutation-based importance provides a measure of global feature magnitude but does not capture directionality or instance-level effects. To obtain faithful and direction-aware explanations for the selected static model, we compute TreeSHAP values for the LightGBM classifier across all test clients. TreeSHAP provides an additive decomposition of each prediction into feature-level contributions, yielding both global and local interpretability while remaining consistent with the model's decision structure.

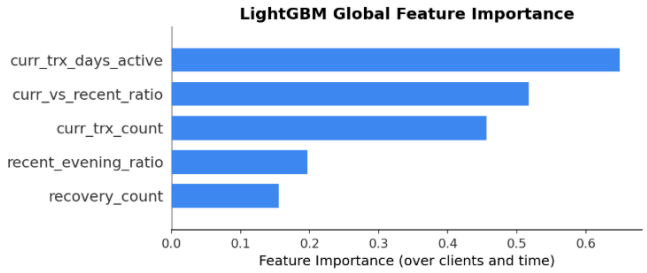

Figure D1: Global Feature Importance

Figure 12 summarizes global feature importance using mean absolute SHAP values aggregated across clients. The most influential features are dominated by **relative and temporally grounded transaction signals**, rather than raw activity levels. In particular, features such as curr_vs_recent_ratio, curr_trx_days_active, and curr_trx_count contribute most strongly to model predictions, indicating that deviations

from recent behavioral baselines are more informative than absolute usage.

Local Attribution: Explaining Why a Specific Client Was Flagged

Effective churn intervention requires understanding not only which clients are at risk, but why a specific prediction is produced. To provide interpretable explanations aligned with the selected modeling approach, we apply TreeSHAP to the LightGBM model, enabling both global and instance-level attribution of churn risk to individual behavioral features.

For a given client c, the predicted churn logit can be expressed as an additive decomposition

$$\hat{y}(x^{(c)}) = \phi_0 + \sum_{i=1}^{d} \phi_i^{(c)},$$

where $\phi_i^{(c)}$ denotes the contribution of feature i to that client's churn probability. This additive decomposition yields a ranked list of the behavioral shifts most responsible for the model's decision.

Figure D2 presents a global summary of SHAP values for the most influential LightGBM features across the test population. The model places greatest emphasis on **relative and temporal transaction signals**, rather than absolute activity levels. In particular, changes in engagement intensity (e.g., curr_trx_days_active, curr_trx_count) and ratios comparing current behavior to recent history (curr_vs_recent_ratio) dominate model decisions. Higher values of these features (shown in red) tend to push predictions toward higher churn risk when they reflect **sharp declines or instability relative to a client's baseline**, while lower values (blue) correspond to more stable engagement patterns.

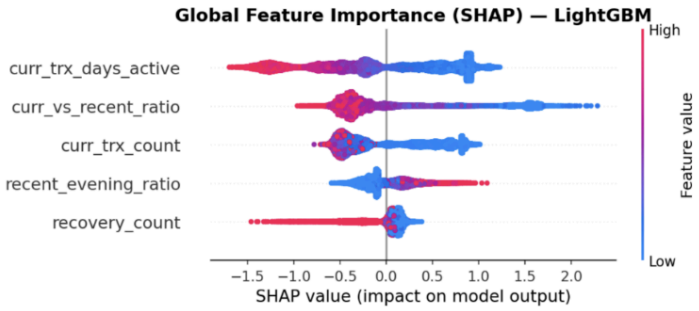

Figure D2: Global Feature Importance (Client-level)

Short-horizon behavioral signals, such as shifts in evening activity and recovery-related features, provide additional context by highlighting recent disruptions following periods of reduced engagement. Importantly, these explanations demonstrate that the model is not reacting to static usage levels, but to **behavioral change and deviation over time**, consistent with the intuition that churn is driven by evolving engagement patterns.

Overall, the LightGBM interpretability analysis confirms that **carefully engineered temporal and ratio-based features capture the key drivers of churn**, enabling transparent, actionable explanations without the added complexity of sequential architectures.

Appendix D: Literature Review

Early customer-churn research in banking relied more on single-modal, structured-data machine learning models, including logistic regression, decision trees, SVMs, Random Forest, and XGBoost. Benchmark studies evaluating these traditional models show that tree-based ensemble methods, especially XGBoost, tend to achieve the strongest performance on banking churn datasets, with reported metrics such as 79% recall and ROC-AUC of 0.87 (Zhang, 2024). Beyond binary

classification, researchers have also explored time-to-event modeling through survival analysis. Random Survival Forests (RSF) provide an advantage by predicting both whether and when churn will occur, outperforming classical survival methods like Cox Proportional Hazards in time-dependent accuracy (Mustonen, 2025).

Despite these advances, traditional churn models remain limited by their dependence on narrow, unimodal feature sets, typically restricted to CRM variables or aggregated transactional indicators. Research on uplift modeling further illustrates this limitation: while uplift frameworks help separate treatment and control effects for retention campaigns, they still depend on structured, single-channel features and do not capture the full behavioral context of customers (Belbahri et al., 2020 – Qini-Based Uplift Regression).

More recent work has shifted toward multimodal churn prediction, motivated by the recognition that customer decisions emerge from complex signals across multiple sources. Rudd et al. propose a multimodal fusion learning framework that integrates customer emotions from voice recordings, financial literacy indicators, and behavioral financial data in their article. Their model significantly outperforms single-source approaches, achieving 91.2% test accuracy through hybrid fusion and demonstrating strong correlations between negative emotions, low financial literacy, and higher churn risk (Rudd et al., 2023 – Multimodal Fusion Learning). This evidence suggests that churn behavior cannot be sufficiently understood from transaction data alone.

At the same time, advances in transaction-sequence modeling illustrate the benefits of deep representation learning for financial behavior. Yugay and Zaytsev show that combining contrastive self-supervised learning with generative masked-event modeling produces embeddings that capture both global customer habits and local event-specific variation in transaction sequences. Their hybrid CMLM-CoLES approach (a sequence-representation learning framework) outperforms single-objective self-supervised models across multiple downstream tasks, indicating that traditional feature engineering often fails to capture the temporal richness of transaction data (Yugay & Zaytsev, 2024 – Contrastive + Generative Learning for Transaction Sequences).

Appendix E: Feature Engineering

1. Feature Engineering

Feature engineering for churn prediction follows strict temporal separation to prevent data leakage. All features are computed from a fixed 6-month historical window ending at the reference month *M*, while churn labels are derived exclusively from future observations. This ensures the model learns predictive patterns rather than retrospective correlations. Full diagnostic procedures and validation results are reported in Appendix A.

2. Window Structure

For each reference month *M*, features are extracted from the window [*M-5, M*], comprising the 6 most recent months of customer activity.

$$\text{Feature Window} = \{M-5, M-4, M-3, M-2, M-1, M\}$$

This fixed-length window ensures temporal consistency across all observations and captures medium-term behavioral patterns sufficient for identifying disengagement trajectories.

Metric	Value
Total Features	62
Training Samples	1,082,840
Test Samples	270,405
Feature Window	6 months
Churn Rate	9.45%

3. Feature Categories

Features are organized into 6 categories based on their sources and temporal scope. We have a total of 62 features.

Category	Count	Description
Current Month	3	Activity snapshot at reference month
Recent Aggregates	8	Aggregated patterns over 6 months
Time-Series	22	Trend, momentum, and volatility
Ratio Features	1	Current vs historical comparison
Dialog Features	12	Customer service interactions
Geographic Features	16	Location-based patterns
Total	**62**	

a. Current Month Features

Current month features capture the customer's activity state at the reference month, providing an immediate snapshot of engagement.

i. Transaction Days Active

The number of unique days with at least one transaction in the reference month.

$$\texttt{curr\_trx\_days\_active}(c, M) = |\{d : \text{customer } c \text{ had transaction on day } d \text{ in month } M\}|$$

This feature ranges from 0 to 31 and measures engagement frequency. A customer active on 20 days exhibits stronger engagement than one active on only 2 days. Sudden drops in this metric often precede churn.

ii. Transaction Count

Total number of transactions in the reference month.

$$\texttt{curr\_trx\_count}(c, M) = \sum_{t \in M} \mathbb{I}[\text{transaction } t \text{ by } c]$$

Combined with days active, this reveals transaction patterns. High count with few active days indicates bursty behavior, while moderate count spread across many days suggests consistent engagement.

b. Recent Aggregate Features

These features aggregate customer behavior across the entire 6-month feature window, capturing longer-term patterns.

i. Months Active

Count of months with at least one transaction.

$$\texttt{recent\_trx\_months\_active}(c, M) = \sum_{m=M-5}^{M} \mathbb{I}[\texttt{trx\_count}(c, m) > 0]$$

This feature ranges from 0 to 6. A value of 6 indicates consistent engagement every month, while values of 1–2 suggest sporadic activity and potential disengagement. This feature is particularly effective for temporal models.

ii. Time-of-Day Ratios

Behavioral patterns are captured through time-of-day transaction ratios.

$$\texttt{recent\_evening\_ratio}(c, M) = \frac{\sum_{m=M-5}^{M} \texttt{evening\_trx}(c, m)}{\sum_{m=M-5}^{M} \texttt{total\_trx}(c, m)}$$

The evening transactions occur between 18:00 and 23:59. Similar ratios are computed for morning (06:00–11:59) and weekend transactions. Changes in these patterns may indicate lifestyle changes affecting churn probability.

c. Ratio Features

i. Current vs Recent Ratio

The primary trend indicator compares current activity to the 6-month average.

$$\texttt{curr\_vs\_recent\_ratio}(c, M) = \frac{\texttt{curr\_trx\_count}(c, M)}{\frac{1}{6}\sum_{m=M-5}^{M} \texttt{trx\_count}(c, m) + \varepsilon}$$

- Ratio < 0.5: Severe decline, strong churn signal
- Ratio ≈ 1.0: Stable activity
- Ratio > 1.2: Growing engagement

This feature directly measures the direction and magnitude of activity change relative to established patterns.

d. Time Series Features

Time-series features capture temporal dynamics including trends, momentum, and volatility across the 6-month window.

i. Trend Slope

A linear regression is fitted to monthly transaction counts.

$$\texttt{trx\_count}(c, m) = \alpha + \beta \cdot m + \epsilon$$

$$\texttt{trend\_count\_slope}(c, M) = \beta$$

Negative slopes indicate declining activity trajectories and are strong predictors of churn.

ii. Coefficient of Variation

Volatility is measured using the coefficient of variation.

$$\texttt{cv\_count}(c, M) = \frac{\sigma(\texttt{counts})}{\mu(\texttt{counts}) + \varepsilon}$$

Where counts = [trx_count(c, m)] for m in {M-5, …, M}. Low values indicate stable, predictable customers, while high values suggest erratic behavior that may precede churn.

iii. Maximum Drawdown

Borrowed from financial analysis, maximum drawdown captures the largest peak-to-trough decline.

$$\texttt{max\_drawdown}(c, M) = \max_{i<j} \frac{\texttt{peak}_i - \texttt{count}_j}{\texttt{peak}_i + \varepsilon}$$

This metric identifies customers who experienced significant activity drops, even if partially recovered.

iv. Recovery Detection

Recovery patterns are detected using V-shaped activity profiles.

$$\texttt{recovery}(c, M) = \begin{cases} \frac{\texttt{count}_M - \texttt{count}_{M-1}}{\texttt{count}_{M-2} - \texttt{count}_{M-1}}, & \text{if } \texttt{count}_M > \texttt{count}_{M-1} < \texttt{count}_{M-2} \\ 0, & \text{otherwise} \end{cases}$$

Customers showing recovery after a dip demonstrate resilience and lower churn risk.

e. Dialog Features

Customer service interactions are captured through dialog features computed over the 6-month window.

i. Dialog Count and Frequency

$$\texttt{dialog\_count}(c, M) = \sum_{m=M-5}^{M} \texttt{dialogs}(c, m)$$

High dialog counts may indicate customer issues or, alternatively, active engagement. Context from other features helps disambiguate interpretation.

ii. Dialog Embeddings

Customer service conversations are represented using pre-trained language model embeddings. The mean embedding across all dialogs is computed, and the first 10 principal dimensions are retained as features.

$$\texttt{dialog\_emb}_i(c, M) = \left(\frac{1}{N} \sum_{d=1}^{N} e_d \right)_i \quad \text{for } i \in \{0, \ldots, 9\}$$

These features capture semantic content of customer interactions without exposing raw text.

f. Geographic Features

Location-based features are derived from geo-tagged transactions.

i. Location Diversity

Unique locations visited at different precision levels.

$$\texttt{geo\_unique}(c, M) = |\{h : h \in \text{geohashes for } c \text{ in } [M-5, M]\}|$$

Higher diversity may indicate travel patterns or mobility, while concentration at few locations suggests routine behavior.

ii. Location Concentration

The proportion of transactions at the primary location.

$$\texttt{geo\_concentration}(c, M) = \frac{\max_h \texttt{count}(h)}{\sum_h \texttt{count}(h)}$$

High concentration indicates a stable home base, while low concentration suggests distributed activity patterns.

Appendix F: Churn Calculation

1. Baseline Average (6 months)

The baseline monthly activity is computed as the average transaction count over months *M−6* to *M−1*.

$$\text{baseline\_avg}(\boldsymbol{c}, \boldsymbol{M}) = \frac{\mathbf{1}}{\mathbf{6}} \sum_{\boldsymbol{m}=\boldsymbol{M-6}}^{\boldsymbol{M-1}} \text{trx}(\boldsymbol{c}, \boldsymbol{m})$$

2. **Expected Future Activity (3 months)**
   Expected activity is defined as 3 times the baseline average.

$$\text{expected\_future}(c, M) = 3 \times \text{baseline\_avg}(c, M)$$

3. **Observed Future Activity (3 months)**
   Observed activity is the sum of transactions in months M+1 to M+3.

$$\text{actual\_future}(c, M) = \sum_{m=M+1}^{M+3} \text{trx}(c, m)$$

4. **Churn Ratio and Label**
   We compute a churn ratio as observed divided by expected activity. We include a small constant ϵ in the denominator to stabilize the churn ratio when expected future activity is very small. This avoids extreme ratios and pathological variance caused by near-zero baselines, which can otherwise disproportionately amplify minor fluctuations in observed activity. In the end, a customer is labeled as churned if this ratio is below a predefined threshold (0.2).

$$\text{churn\_ratio}(c, M) = \frac{\text{actual\_future}(c, M)}{\text{expected\_future}(c, M) + \varepsilon}$$

$$\text{churned}(c, M) = \begin{cases} 1, & \text{if churn\_ratio}(c, M) < 0.20 \\ 0, & \text{otherwise} \end{cases}$$

Appendix G: Train-Test Split Strategy

We partition the data using **reference-month-based splits**, consistent with the churn label construction described in Section 2.3.2.

Each churn label requires nine months of data per reference month: six months to compute the baseline average transaction count and three months to observe future activity for churn ratio computation. As a result, **May 2022** is the earliest reference month for which a valid churn label can be constructed given dataset coverage; earlier months lack sufficient historical context and are excluded from modeling.

Under this constraint, the training set includes observations with reference months **2022-05 to 2022-07**, while the test set includes **2022-08 and 2022-09**. For each reference month, all features are computed strictly using information available up to that month, and churn labels are derived from activity in the subsequent three-month window. Churn prevalence is similar across splits (Train: 9.25%, Test: 9.84%).