Multimodal Banking (MDB) Attrition Project

Data Science Institute
Columbia University

Kshitij Nitin Patil, Debbie Dai, Saravanan Hari Baskaran, Xinyan Li, Yuntong Ju
Tianyue Wang, Young Cheong
Adam Kelleher

TD Bank

1. Abstraction and Business Problem

Customer attrition poses a significant financial risk to the banking industry, where even minor decreases in retention can result in millions of dollars in annual lost revenue. U.S **retail-bank retention** fell from **78% to 76%** in 2023, highlighting an urgent need for more effective prediction and intervention strategies.

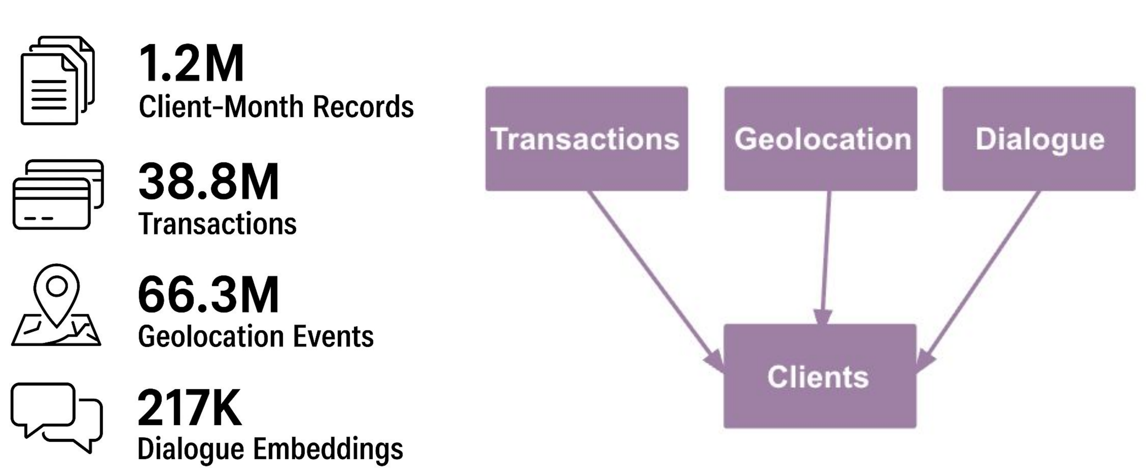
Most existing churn models rely on **single-modal**, structured data, failing to capture the rich, early warning signals hidden within customers' **cross-channel behaviors**. This project addresses this gap by developing a multimodal predictive framework that integrates diverse data streams to identify at-risk clients earlier and more accurately

Project Goal:
To build and validate a machine learning pipeline that fuses transactional, geographic and dialogue data to predict customer churn, enabling proactive and personalized retention efforts.

2. Dataset: MDB-mini

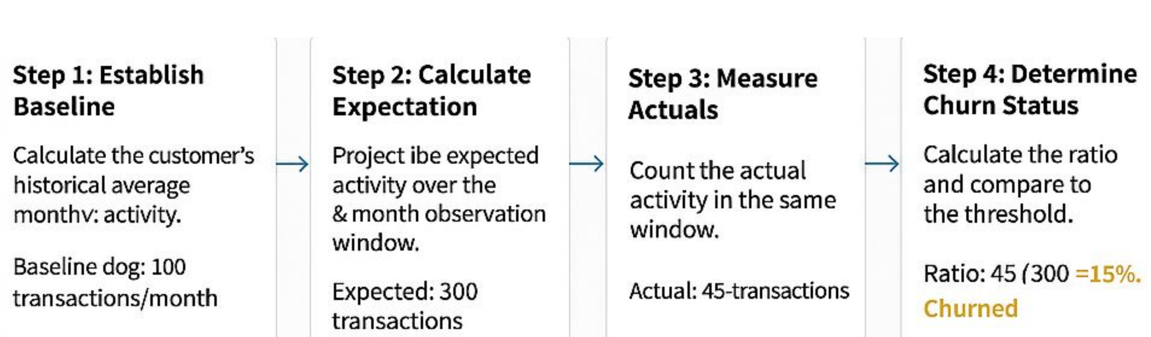
The project uses the Multimodal Banking Dataset (MBD) published by the Sber AI Lab; for our work, we focus on the mini version of this dataset.

Data Scale & Modalities:



3. Target Definition

A customer is defined as **churned** when their activity drops below **20%** below of their expected level. This threshold is the core of our target variable.



4. Methodology: A Multimodal Framework

Our approach follows a structured pipeline from raw data integration to sophisticated modeling, designed to extract deep behavioral patterns.

4.1 Feature Engineering

Our feature engineering framework captures a holistic, multi-faceted view of customer behavior across activity, engagement, and geospatial patterns. All features are derived strictly within the 3-month feature window to prevent data leakage.

Key Feature Groups:

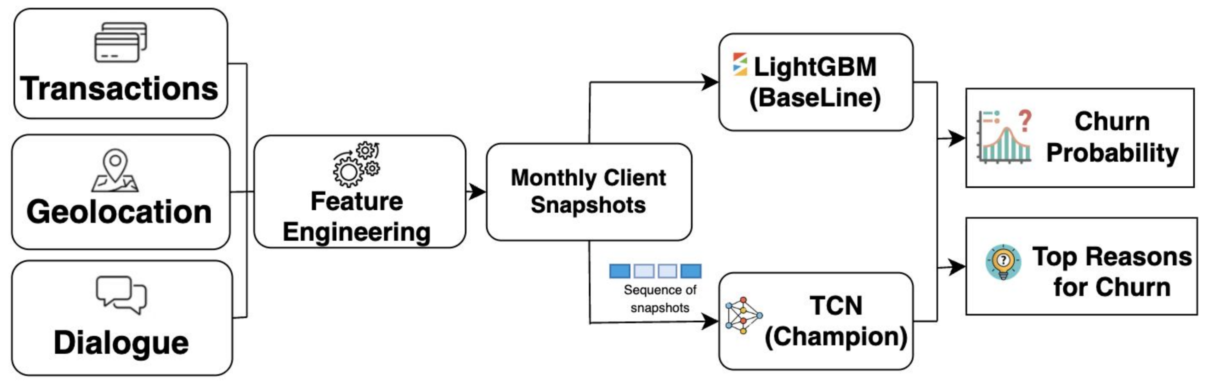
- **Current Month (3 features):** Snapshot of most recent activity (counts, spend, intensity).
- **Recent 3 Months (8 features):** Short-term behavioral patterns through aggregated trends.
- **Time-Series Dynamics (22 features):** Momentum, volatility, stability, and temporal shifts in behavior.
- **Dialog / Customer Service (12 features):** Frequency, recency, and intensity of interaction with support channels.
- **Geospatial Behavior (17 features):** Movement diversity, location dominance, entropy, and stability of top locations.

4.2 Model Architecture

We evaluated two classes of models to establish a robust performance benchmark and capture temporal dynamics.

A. Baseline Model: LightGBM: A powerful gradient-boosted model chosen for its high performance on heterogeneous tabular data, speed, and inherent handling of class imbalance. This established a strong, fair baseline.

B. Champion Model: Temporal Convolutional Network (TCN): A sequential deep learning model selected to analyze client behavior over time. By processing sequences of monthly features, the TCN can identify subtle, time-dependent patterns and deteriorating engagement trends that static models often miss.



5. Results & Key Findings

LightGBM
Train: 202,357 samples (9.25% churn) | Test: 34,160 samples (9.84% churn)
Features: 71 | Positive Class Weight: 9.81 | Optimal Threshold: 0.337

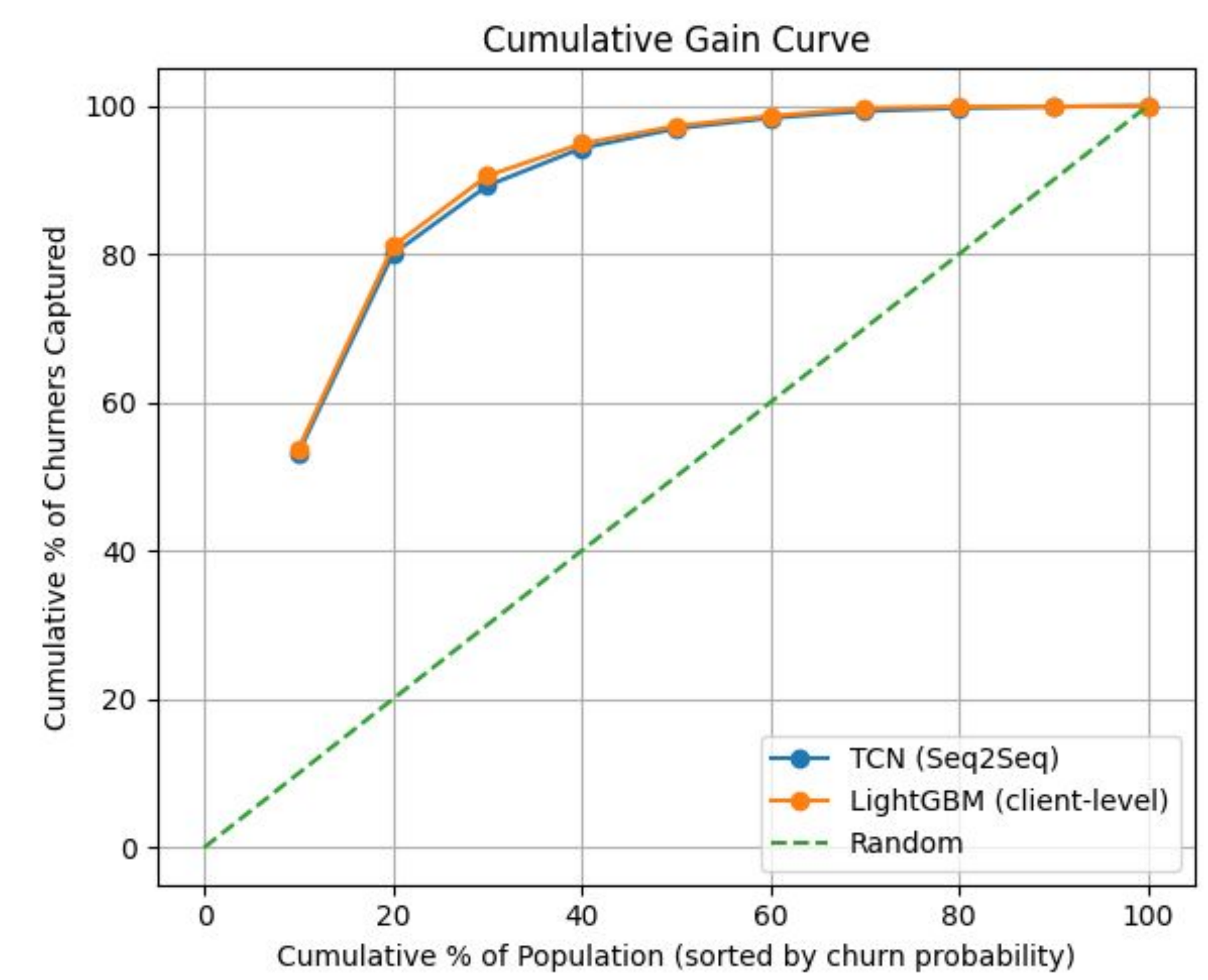
TCN
Train: 202,357 samples (9.25% churn) | Test: 34,160 samples (9.84% churn)
Features: 71 | Optimal Threshold: 0.497

Metric	LightGBM	TCN	Difference	Winner
AUC-ROC	0.9138	0.9109	-0.0029	LightGBM
Precision	0.5025	0.5619	+0.0594	TCN (+11.8%)
Recall	0.6830	0.6776	-0.0054	LightGBM
F1 Score	0.5790	0.6143	+0.0353	TCN (+6.1%)

When comparing, TCN (Seq2Seq) performs comparably to LightGBM in ROC-AUC, but delivers stronger precision, F1-score, and PR-AUC which are metrics that matter most in churn prediction. This indicates that **TCN is better at predicting the true churners while reducing false positives**. It is more effective on decision making focusing on retention.

Evaluation: Cumulative Gain curve

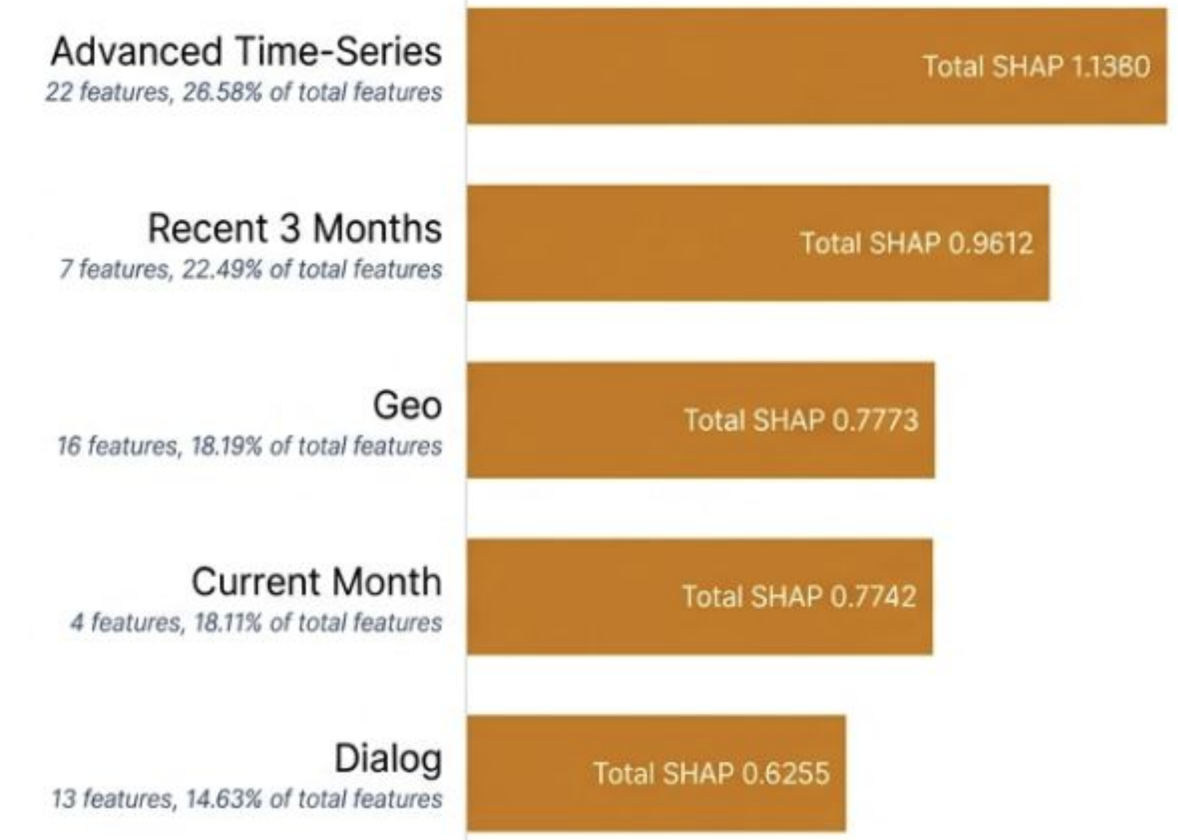
The Cumulative Gain Curve demonstrates the real business impact of our models by showing how many actual churners we can capture when we target a given percentage of the highest-risk clients.



Key Finding: The sequence-based TCN model consistently outperforms client-level LightGBM by a small margin, capturing **~80% of churners** by targeting only the **top 20% at-risk clients.**

6. SHAP Analysis

The Four Dimensions of Churn Risk



- **Advanced Time-Series** features contribute the most, indicating that patterns such as volatility, trend changes, and temporal irregularities are the strongest signals of churn.
- **Patterns over time** matter more than single events, and TCN can capture when habits break.
- **Geo + dialogue signals** refine predictions but play secondary roles.

7. Conclusion & Business Impact

In General, **the TCN model delivers high accuracy and clear explanations**, capturing behavioral shifts over time that static models miss. In enabled early detection of disengagement by identifying the drops in clident's interaction behaviors. The model transforms churn prediction from a black-box score into transparent, actionable insights.

Business Impact

- **Proactive intervention:** Shifts from reactive win-backs to proactive retention, giving a 1–2 month action window to prevent revenue loss.
- **Targeted ROI:** Focusing on the top 20% highest-risk clients captures 80%+ of likely churners, boosting marketing efficiency.
- **Personalized Offers:** Feature importance allows for hyper-personalized interventions for each specific user.

Future Scope

Expand Data Coverage	Real Time Deployment	A/B Test Interventions
Add newest monthly data & signals	Build live scoring pipeline	Measure lift vs. current strategy

Patel, V. (2024, September 25). *Creating a loyal experience for customers in banking*. Capgemini. Retrieved October 24, 2025.
Belbahri, M., Murua, A., Gandouet, O., & Partovi Nia, V. (2020). *Qini-based uplift regression*. arXiv. https://arxiv.org/abs/1911.12474
Mustonen, V. (2025). *Customer churn estimation in the banking sector with random survival forest* (Master's thesis). Tampere University.
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.