USER CHURN PROJECT | ML Model Results

Prepared for the Waze Leadership Team

waze

Issue / Problem

Waze's data team is building a predictive model to identify users at risk of monthly churn (uninstalling or ceasing app use). This milestone delivered the first production-ready classifier, tested on held-out data and paired with a business-aligned decision threshold so retention teams can act with confidence.

This report summarizes Milestone 6 and its impact for any future development.

Response

- We trained and compared two ensemble classifiers—**Random Forest** and **XGBoost**—using a three-way split (train/validation/test). A separate validation set enabled objective model selection; only the final champion was evaluated once on the untouched test set to estimate future performance.
- We **optimized first for recall** (to minimize missed churners) while tracking **precision**, **ROC AUC**, and **PR AUC**.
- We then selected an operational decision threshold from the Precision–Recall curve to align with Waze's outreach strategy.

Key Insights

Champion model: XGBoost outperformed Random Forest on recall.

- **Validation: Recall ≈ 0.65**, **Precision ≈ 0.34**, **ROC-AUC ≈ 0.75** (stable across folds).
- **Test (default 0.50 cutoff): Recall ≈ 0.61**, **Precision ≈ 0.31**, **F1 ≈ 0.41**, **Accuracy ≈ 0.69**, **ROC-AUC ≈ 0.72**, **PR-AUC (AP) ≈ 0.354**.
- **Class imbalance:** Positives ≈ **18%** (baseline AP ≈ **0.175**). Our AP ≈ **0.354** is ~**2x** baseline.
- **Lift vs. dummy:** Baseline recall ≈ **0.16;** model delivers **~0.52–0.61** (≈ **3–4x** more churners captured).

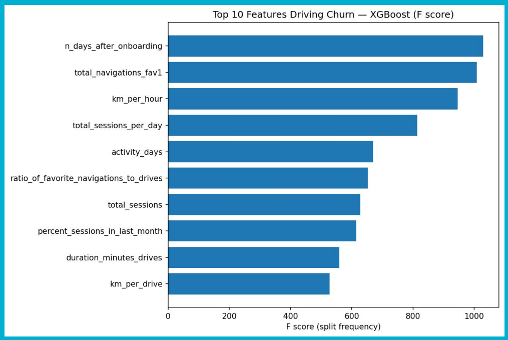
Operational threshold: To meet a **≥50% recall** production target, we apply a **bootstrap-conservative** policy on validation (1st percentile), yielding a **frozen threshold ≈ 0.575**. On the unseen test set this delivers **~0.52 recall** at **~0.34 precision** (**F1 ≈ 0.41**), **accuracy ≈ 0.74**, and **flag rate ≈ 26.7%**. The conservative policy controls outreach volume while meeting the recall target.

Behavioral drivers:

- **Early tenure risk:** The first **60–90 days** after onboarding are most fragile.
- **Usage intensity / recency:** Lower recent activity and session intensity increase churn risk.

Impact

- **Retention leverage:** With the **validation-frozen threshold (~0.575)**, the model captures **~52%** of churners at **~34%** precision (flag rate **~26.7%**). If capacity allows, operating at **0.50** increases recall to **~61%** (precision **~31%**) with more outreach.
- **Transparency & actionability:** Confusion matrices quantify FPs/FNs; **feature importance (F-score / split frequency)**explains why users are flagged, guiding onboarding and re-engagement tactics.
- **Rigor & cost control:** Thresholds are **selected on validation via bootstrap** (no test peeking). **Model + policy** are saved for reproducibility. **Risk tiers** support capacity planning.



Recommendations

➔ **Deploy champion + policy:** Persist the fitted **XGBoost** and the **frozen threshold ≈ 0.575**; score on a **daily/weekly** cadence.

➔ **Onboarding focus:** Invest in interventions during the **first 60–90 days** (guided tutorials, timely nudges, first-week habit formation).

➔ **Measure & tune:** Track **precision, recall, flag rate**, and **campaign ROI**; monitor drift and **retrain on a schedule** or when drift triggers fire.

➔ **Data enrichment (next iteration):** Add richer **recency/velocity** features, in-app **notification/response** signals, and session-level patterns to **raise precision at similar recall**.