Predicting User Churn

Anoosh Moutafian 2024

waze

Waze provides satellite navigation software on smartphones

- Revenue generated via ad sales
- App relies on crowdsourced information
- 150 million monthly active users

Dataset: Synthetic dataset constructed by Waze

- 14,299 complete rows
- 13 features (most cover one-month time period)
- Data labeled churn/retain (reflects user behavior behavior during one-month period)

Problem Statement

User churn during dataset month

- 18%
- 2536 users churn /14299 total

Goal: Reduce user churn by 10%

- 18% → 16.2%
- 2536 → 2316 users churn
- 220 users retained/14299 total

Questions:

How can we predict user churn?

Which features contribute most strongly?

Results of successful retentions:

- Increased ad impressions at $.002 each
- Increased app quality with additional crowdsourced info

Data Cleaning, Exploration, Wrangling and Preprocessing

Of the original 14,999 observations (each representing a Waze user), 700 were removed due to missing churn information

Features were explored individually and in relationship to other features

The imbalanced churn/retain feature was balanced using random undersampling
18% $\longrightarrow$ 50% churn rate

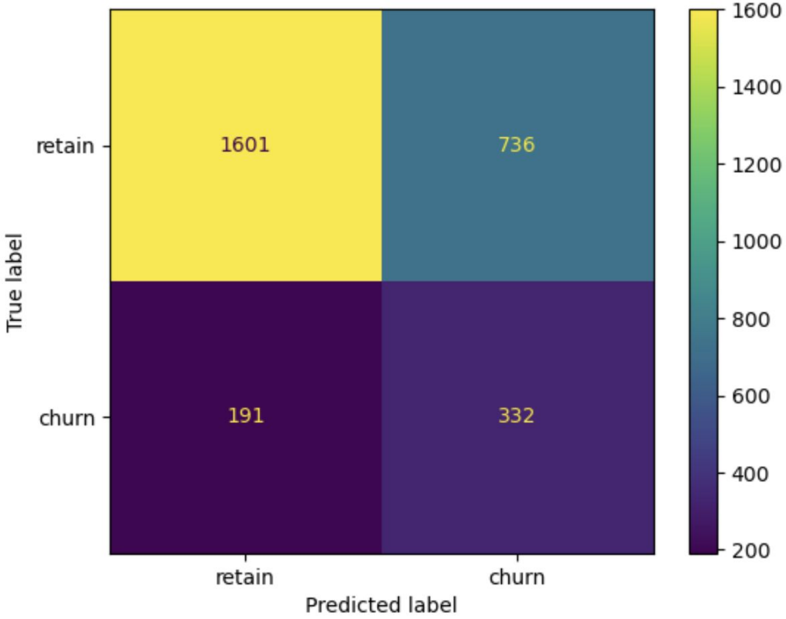
The data was split into 80% train and 20% test sets

Dummies were made from the single independent categorical feature

Numerical features were standardized

Modeling: Gradient Boost using XGBoost was the most successful

	precision	recall	f1-score	support
0	0.90	0.63	0.74	2337
1	0.29	0.67	0.41	523
accuracy			0.64	2860
macro avg	0.59	0.65	0.57	2860
weighted avg	0.79	0.64	0.68	2860



Relative feature importance as identified by XGB

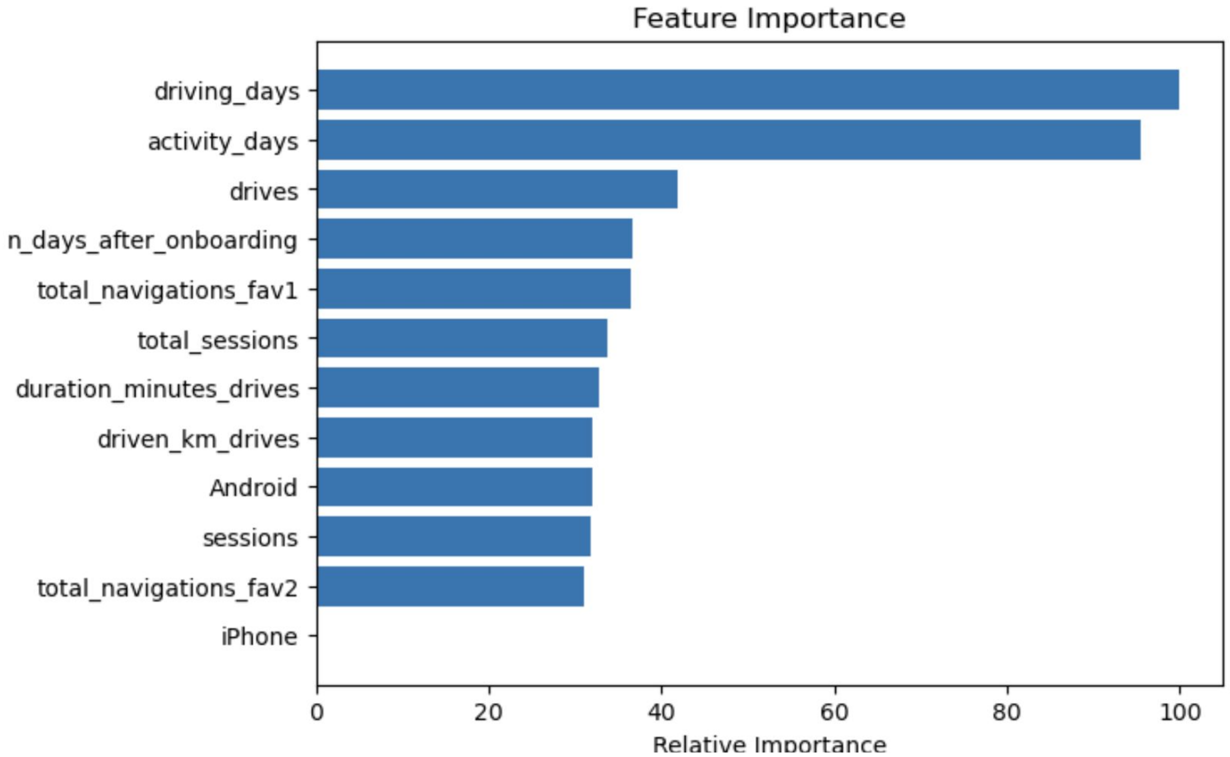
Best predictors of user churn:

driving_days:

Number of days the user drives (at least 1 km) during the month

activity_days:

Number of days the user opens the app during the month



Next Steps:

Tune gradient boost model hyperparameters

Convert data to DMatrix (XGBoost proprietary data structure)

Experiment with class balancing techniques