Final Report: Waze user churn
Anoosh Moutafian 2024

Problem Statement

Waze, released in 2006, is a subsidiary company of Google that provides satellite navigation software. Free to download, the Waze app generates revenue through ad sales. As the app relies on crowdsourced information, keeping user numbers up is especially critical. With 140 million monthly active users (as of 2022), moving the needle on churn rates even a few percentage points has large implications.

While user behavior is ultimately out of company control, this investigation aims to identify features associated with churn and suggest steps to reduce user churn by 10%. Of the 14,299 customers, 2536 (18%) churned. The goal is to reduce churn by 10%. Therefore, a successful outcome would be if during a similar period and population, only 16.2%, or 2316 customers churned.

Dataset and Data Wrangling

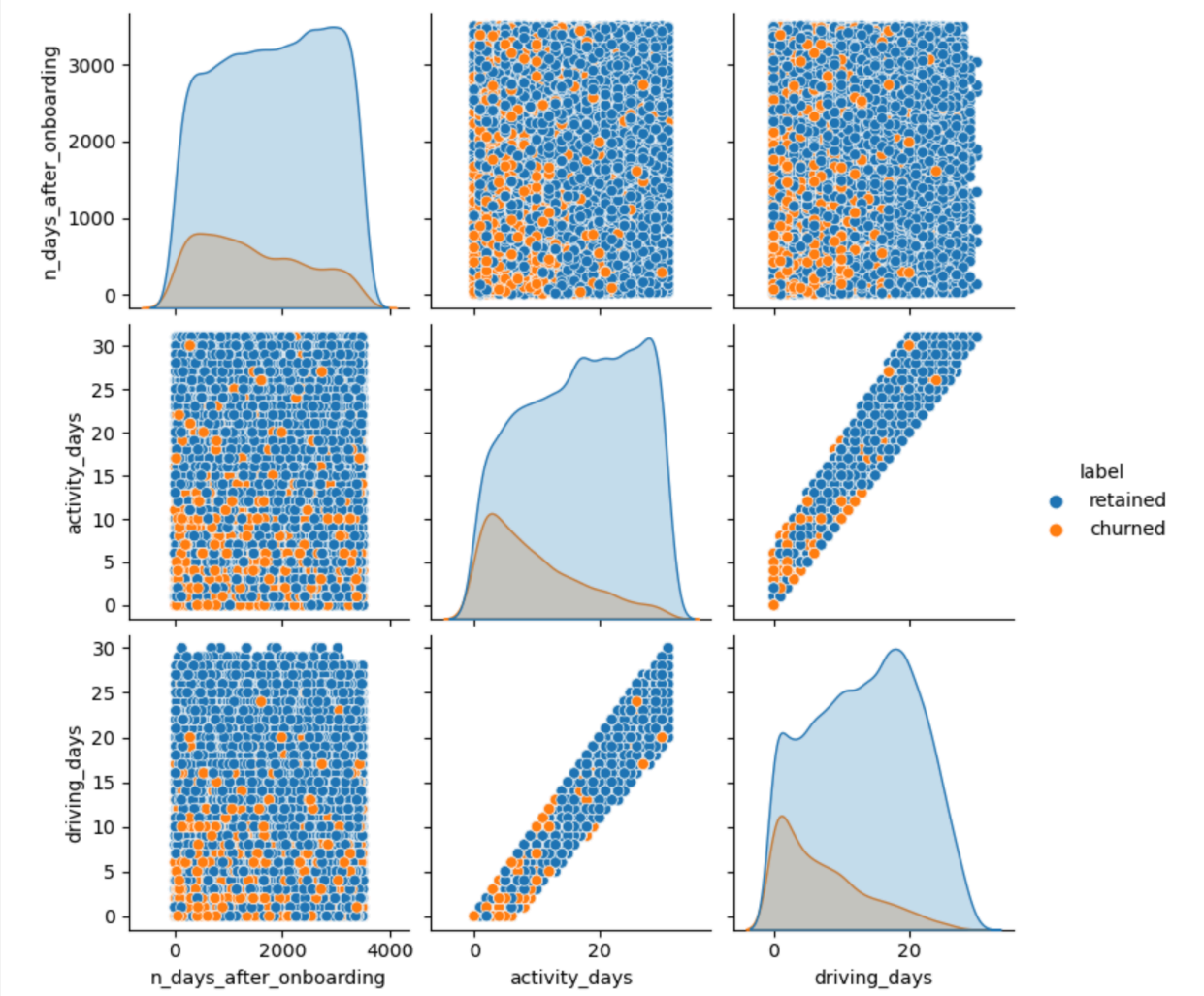
A synthetic dataset constructed by Waze with 13 features was used as the basis of the analysis. Of the 14,999 original rows, 700 had to be discarded due to missing churn data. Remaining rows were complete with no nulls or duplicates with reasonable value ranges. Most columns covered the time period of one month. The churn/remain label reflects user behavior during this one-month period.
https://github.com/amoutafian/waze_capstone/blob/main/data/waze_dataset.csv

Exploratory Data Analysis

The features with the strongest relationship to the dependent variable seem to be “activity_days,” “driving_days" and “n_days_after_onboarding”.

"activity_days" : Number of days the user opens the app during the month
"driving_days" : Number of days the user drives (at least 1 km) during the month
“n_days_after_onboarding”: Number of days since first app open

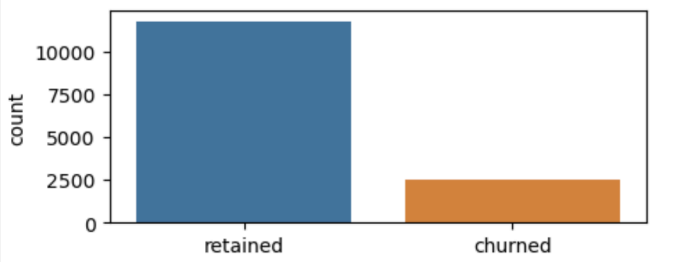
The 'n_days_after_onboarding' is particularly valuable as it represents one of the few features with information about the user's history. The shortest user tenure is only 4 days while the longest is around 9.5 years.

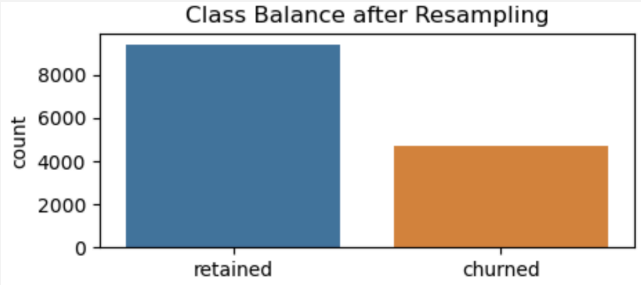


Dataset Limitations: This is not a time series, we have information about what happened during the month, but not when it happened. Is it that light users are more likely to churn, or is it that users who churn don't have a chance to use heavily as they have left the service mid-data collection?

Pre-Processing

The data is unbalanced. The churn/retention ratio is 2536/14299, or 18%. Random oversampling with Imblearn was applied to balance the classes to 33% churn rate.





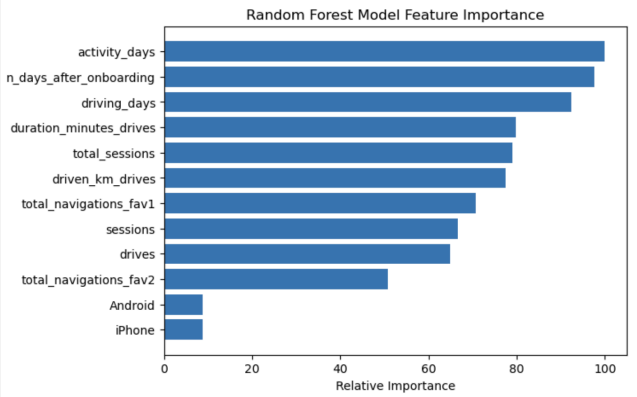
Dummy features were created for the single independent categorical feature ‘device’.

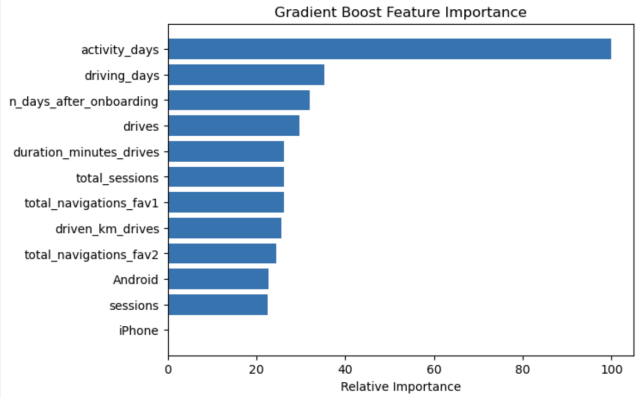
The data was split into testing and training datasets with a test size of 20%.

Numeric features were standardized using Standard Scaler.

Modeling
Logistic Regression, Random Forest, K Nearest Neighbors and Gradient Boost (XGBoost) models were used.

Feature importance: As expected, driving_days and activity_days were the most important features according to the gradient boost models. The random forest model ranked n_days after onboarding as second most important and activity days, days after onboarding and driving days as highly important, while the gradient boost model dropped relative importance to under 40% for the second ranked feature.





Conclusion:

Logistic Regression:

	precision	recall	f1-score	support
0	0.89	0.77	0.83	2337
1	0.36	0.58	0.45	523
accuracy			0.74	2860
macro avg	0.63	0.68	0.64	2860
weighted avg	0.80	0.74	0.76	2860

Random Forest:

	precision	recall	f1-score	support
0	0.84	0.95	0.89	2337
1	0.48	0.19	0.27	523
accuracy			0.81	2860
macro avg	0.66	0.57	0.58	2860
weighted avg	0.77	0.81	0.78	2860

Gradient Boost:

	precision	recall	f1-score	support
0	0.85	0.89	0.87	2337
1	0.39	0.32	0.35	523
accuracy			0.78	2860
macro avg	0.62	0.60	0.61	2860
weighted avg	0.77	0.78	0.78	2860

Even without optimization, the gradient boost model was competitive with the second highest accuracy score. As our primary business interest is to identify churn, we may want to prioritize sensitivity (recall), the true positive rate.

It would be worthwhile to discuss how the model results will be used with stakeholders. If the planned intervention on customers at risk of churn is expensive or limited, elevating the importance of precision may be warranted.

Random Forest had the best score for accuracy, but Logistic regression was the most successful in identifying true positives for churn (recall, 1).

Next steps

The gradient boost model could be optimized by tuning hyperparameters and converting data to DMatrix, an XGBoost unique data structure that can improve performance and efficiency.

Class balancing techniques could also be explored. Random oversampling with bootstrapping with a sampling strategy of 0.5 was used which drastically improved model importance, but more experimentation could yield better results.