Predicting User Churn | Regression Modeling Insights

Prepared for: Waze Leadership Team

By Rebecca Iglesias-Flores

ISSUE / PROBLEM

- Waze is experiencing a decline in active users, which affects ad revenue and funding for our unique crowdsourced development business model.
- More resources need to be dedicated to studying user behaviors that prevent user churn. This is part of a larger effort at Waze to increase growth.
- At this point in the project a regression model as been created giving initial insights into those behaviors that indicate user satisfaction and retention.

IMPACT

- Churn reduces Waze's active user base, affects advertising revenue, and impacts data accuracy for navigation insights. A 5% increase in retention could significantly boost long-term engagement.

RECOMMENDATIONS

- Due to low recall, this model should not be used; however, it did prove insightful insofar as it revealed a great need for feature engineering.
- Engineered feature professional_driver, was the 2nd most informative feature.
- Highly correlated features might have contributed to the loss in performance and more curation needs to be done.
- More headway needs to be made in data collection to include drive-level information for each user (Ex. drive times, geographic locations, etc.) and user-specific interactions with the app (Ex. num reports submitted).

RESPONSE

- We developed a multiple linear regression model to predict churn likelihood based on user behavior metrics (e.g., trip frequency, time spent on the app, navigation habits).
- The model was evaluated using a correlation matrix, confusion matrix, and classification report.
- The model has decent precision but very low recall, which means that it makes a lot of false negative predictions and fails to capture users who will churn (see Figure 1)..

KEY INSIGHTS

Figure 1. *Classification Report*

	precision	recall	f1-score	support
retained	0.83	0.98	0.90	2941
churned	0.52	0.09	0.16	634
accuracy			0.82	3575
macro avg	0.68	0.54	0.53	3575
weighted avg	0.78	0.82	0.77	3575

The report indicates poor performance in the ability to recall 9% and make precise predictions 52% of the time for churned users. By contrast, the model is excellent in identifying retained users.

The orange-to-red colors in Figure 2 demonstrate high correlation between features. In a logistic regression model, any high correlation between 2 or more independent features has the potential to negatively affect the efficacy in our model. Although driven_km_drives and duration_minutes_drives were high on the the correlation matrix, activity days was the highest impact feature to predicting user churn. As the number of activity days increases the less likely our user will churn which is what we want!

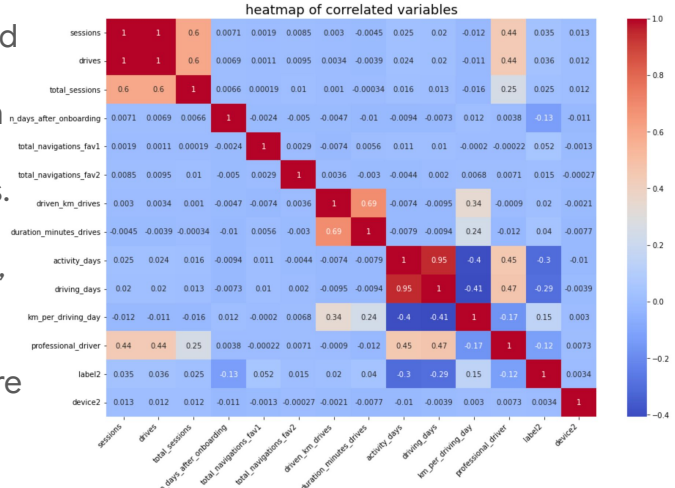

Figure 2.*Heatmap shows correlation between features, high correlation is 0.7.*