Waze Churn Prediction Using Logistic Regression Data Summary

by Michelle Aguiar

Overview

This analysis supports Waze's retention efforts by developing a predictive model to identify users likely to churn. Using the cleaned and engineered dataset from the exploratory phase, the goal was to build an interpretable logistic regression model that highlights key behavioral indicators of churn and evaluates prediction performance.

Project

This project focuses on the modeling phase of churn analysis. Key tasks included:

- Selecting relevant variables after multicollinearity checks
- Encoding categorical variables and scaling numerical features
- Training a logistic regression classifier to predict churn
- Validating assumptions (logit linearity) and interpreting coefficients
- Evaluating model performance using accuracy, precision, recall, F1-score, and confusion matrix

All tasks were conducted in Python using pandas, scikit-learn, and visualization libraries within a Jupyter Notebook environment.

Key Insights

- **Model performance:** The model reached an accuracy of 82.4%, but recall for churned users was low (9%), indicating most at-risk users were not captured by this baseline approach.
- **Driver identification:** The most predictive feature was being a "professional driver," defined as a user with ≥60 drives and ≥15 active driving days. These users had the lowest churn rates.
- **Usage consistency matters:** Features tied to frequency (e.g., activity days) had more predictive value than total usage (e.g., distance or time spent driving).
- **Device type remains irrelevant:** Consistent with the exploratory findings, Android and iPhone users showed no meaningful differences in churn behavior.

Next Steps

- Improve recall by addressing class imbalance using resampling or cost-sensitive methods
- Test tree-based models (e.g., random forest, gradient boosting) for nonlinear signal detection
- Expand feature set to include engagement recency, referral source, and user feedback
- Combine this model with segmentation or clustering to support targeted retention strategies