User Churn Project | Machine Learning Model Results

Prepared for: Waze Leadership Team

ISSUE / PROBLEM

The Waze data team is developing a project to reduce monthly user churn and support overall growth. Churn is defined as users who uninstall or stop using the app. **The main objective is to build a machine learning model to predict user churn.** Predicting the target variable allows us to identify at-risk users before they leave, enabling timely interventions such as incentives or promotions. **This report summarizes key insights from Milestone 6 that may inform future project development and model improvement.**

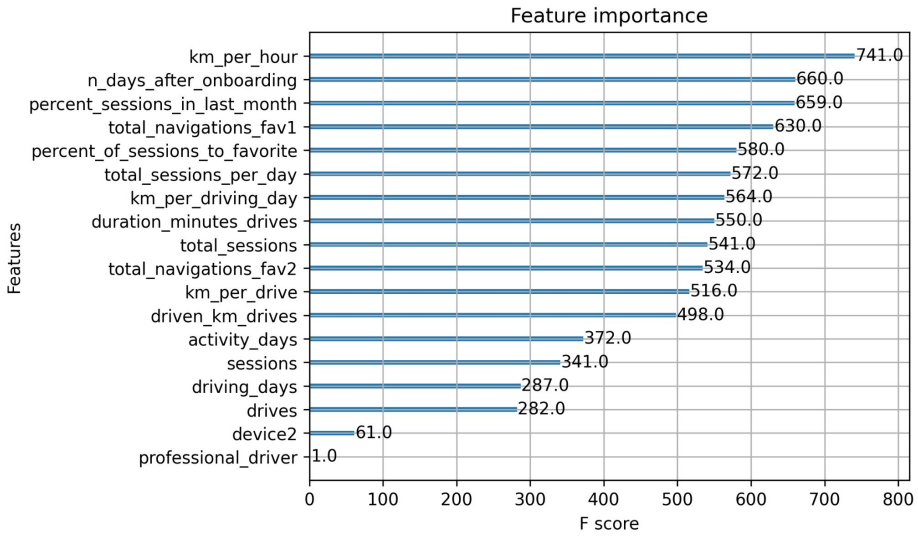
IMPACT

- Milestone 6 models show that **existing data is insufficient to reliably predict user churn,** limiting the model's ability to capture all relevant patterns.
- **More detailed data could improve the model, such as drive-level information (e.g., locations) and user interaction patterns,** like how often they report traffic conditions. This would help the model better capture patterns that indicate potential churn.
- Engineered features are valuable for enhancing model performance. **A second iteration of the User Churn Project is recommended, focusing on richer data collection and targeted feature engineering to improve predictive power.**

RESPONSE

- **To achieve maximum predictive power, the Waze data team developed two models: Random Forest and XGBoost, for cross-comparison.**
- The data was split into training, validation, and test sets. While splitting three ways reduces the data available for training, **using a separate validation set allows for unbiased selection of the champion model and provides a more reliable estimate of future performance than a simple two-way split.** This setup also helps monitor for overfitting and ensures the model generalizes well to unseen data.

KEY INSIGHTS



Engineered features made up 5 of the top 10 most important features: km_per_hour, percent_sessions_in_last_month, percent_of_sessions_to_favorite, total_sessions_per_day, and km_per_driving_day.

The XGBoost model fit the data better than the Random Forest model. Notably, its recall score of 15.9% is nearly double that of the logistic regression model from Milestone 5, while maintaining similar accuracy and precision.

Tree-based model ensembles developed in this project outperformed a single logistic regression model, achieving higher scores across all metrics and requiring less data preprocessing, though they are more complex to interpret.