# Understanding Waze User Churn | ML Model Results

Prepared for: Waze Leadership Team

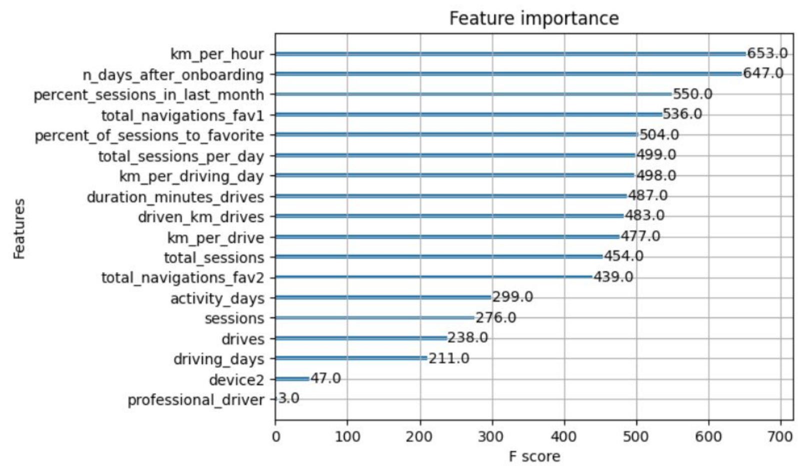
## Project Overview

The Waze data team is currently developing a data analytics project aimed at increasing overall growth by preventing monthly user churn (users who have uninstalled/stopped using) on the Waze app.. The ultimate goal for this project is to develop a machine learning (ML) model that predicts user churn. **This report offers details and key insights from Milestone 6, which could impact the future development of the project, should further work be undertaken.**

## Key Insights

- To obtain a model with the **highest predictive power**, the Waze data team developed two different models to cross-compare results: **random forest** and **XGBoost**.
- To prepare for this work, the data was split into **training, validation, and test sets**. Splitting the data three ways means that there is **less data available** to train the model than splitting just two ways. However, **performing model selection on a separate validation set enables testing of the champion model by itself on the test set, which gives a better estimate of future performance than splitting the data two ways and selecting a champion model by performance on the test data.**
- The ensembles of **tree-based models** in this project milestone are **more valuable** than a **singular logistic regression model** because they achieve **higher scores across all evaluation metrics** and require **less preprocessing of the data.** However, it is **more difficult** to understand how they make their predictions.

## Details




- **Engineered features** accounted for **six of the top 10** features: km_per_hour, percent_sessions_in_last_month, total_sessions_per_day, percent_of_drives_to_favorite, km_per_drive, km_per_driving_day.
- The **XGBoost model fit the data better** than the random forest model. Additionally, it's important to call out that **the recall score (18%)** is nearly **double the score** from the previous **logistic regression model** built in Milestone 5, while still **maintaining a similar accuracy and precision score**.

## Next Steps

- ★ This modeling effort confirms that **the current data is insufficient to consistently predict churn**.
- ★ It would be helpful to have **drive-level information** for each user (such as drive times, geographic locations, etc.). It would probably also be helpful to have **more granular data** to know how users interact with the app i.e. how often do drivers report or confirm road hazard alerts? Finally, it could be helpful to **know the monthly count of unique starting and ending locations** each driver inputs.
- ★ **We recommend gathering further data** as our models demonstrate a critical need for additional data in order to more accurately predict user churn.