Waze Churn Project | Preliminary Data summary

Prepared for: Waze leadership team

ISSUE / PROBLEM

The Waze data team is now working on a data analytics initiative targeted at boosting overall growth by reducing monthly user attrition on the Waze app. For the purposes of this project, churn refers to the number of users who uninstalled or ceased using the Waze app. The ultimate goal of this research is to create an ML model that forecasts customer attrition. This report provides facts and significant insights from Milestone 6, which may have an impact on the project's future development if additional work is completed.

IMPACT

➔ The ML models developed for Milestone 6 show an urgent need for more data to better anticipate user churn.

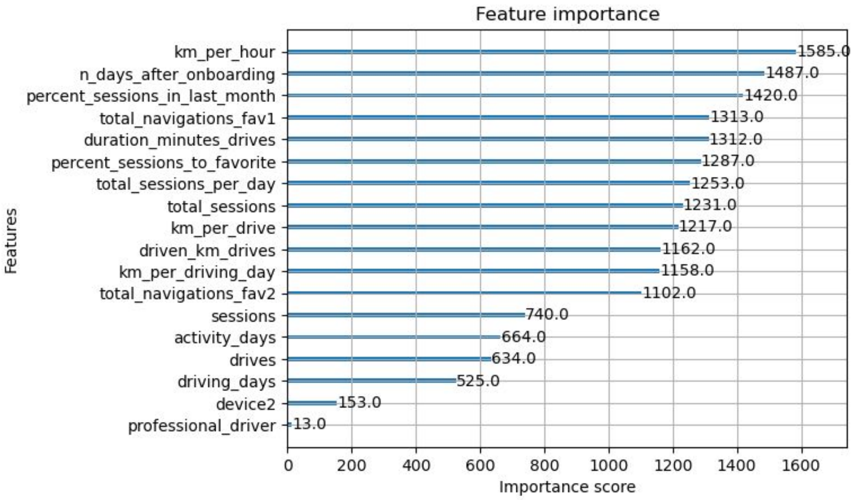
➔ This modeling attempt demonstrates that current data is insufficient to reliably predict churn. It would be useful to include drive-level information for each user (e.g., travel times, geographic locations). It would also be beneficial to have more specific data to understand how users interact with the app. For example, how frequently do they report and confirm road hazard alerts? Finally, knowing the monthly count of each driver's unique starting and ending places could be useful.

➔ Because engineered features have proven to be an effective strategy for enhancing ML model performance, the Waze team suggests a second iteration of the User Churn Project.

RESPONSE

- To achieve the maximum prediction potential, the Waze data team created two models to compare results: random forest and XGBoost.
- To prepare for this job, the data was divided into training, validation, and test sets. Splitting the data three times reduces the amount of data available to train the model compared to splitting it twice. However, executing model selection on a separate validation set allows for independent testing of the champion model on the test set, which provides a better forecast of future performance than splitting the data two ways and selecting a champion model based on test data performance.

KEY INSIGHTS



- Six of the top ten attributes were engineered, including km_per_hour, percent_sessions_in_last_month, total_sessions_per_day, percent_of_drives_to_favorite, km_per_drive, and km_per_driving_day.
- The XGBoost model fit the data more accurately than the random forest model. Furthermore, it is worth noting that the recall score (17%) is approximately double that of the prior logistic regression model created in Milestone 5, while keeping identical accuracy and precision scores.
- In this project milestone, ensembles of tree-based models outperform a single logistic regression model because they get higher scores across all assessment measures and require less data preprocessing. However, it is more difficult to comprehend how they create their forecasts.