User Churn Project | ML Model Results

Prepared for: Waze Leadership Team

ISSUE / PROBLEM

In this project, the Waze data team is working on a data analytics solution to reduce the monthly user churn on the Waze app. User churn refers to the users who have either uninstalled the Waze app or stopped using it. The main objective of this project is to create a machine learning (ML) model that can predict user churn. **This report presents the findings and key insights from Milestone 6, which may influence the future direction of the project, if it is continued.**

IMPACT

➔ **Milestone 6 ML models show that more data is needed to predict user churn better.**

➔ **Current data is not enough to capture churn patterns reliably.**

➔ **Drive-level data (e.g., drive times, locations) would be useful to have.**

➔ **App-level data (e.g., road hazard alerts frequency) would be informative to have.**

➔ **Monthly count of unique start and end locations per driver would be helpful to have.**

➔ **Engineered features can boost ML model performance, so a second iteration of User Churn Project is suggested.**

RESPONSE

- The Waze data team used two models to compare their predictive power: random forest and XGBoost.
- The data was divided into three sets: training, validation, and test. This reduces the amount of data for training, but **allows for model selection on a validation set and testing on a test set, which gives a more accurate estimate of future performance than using two sets and testing on the same data used for selection**.

KEY INSIGHTS

Feature importance
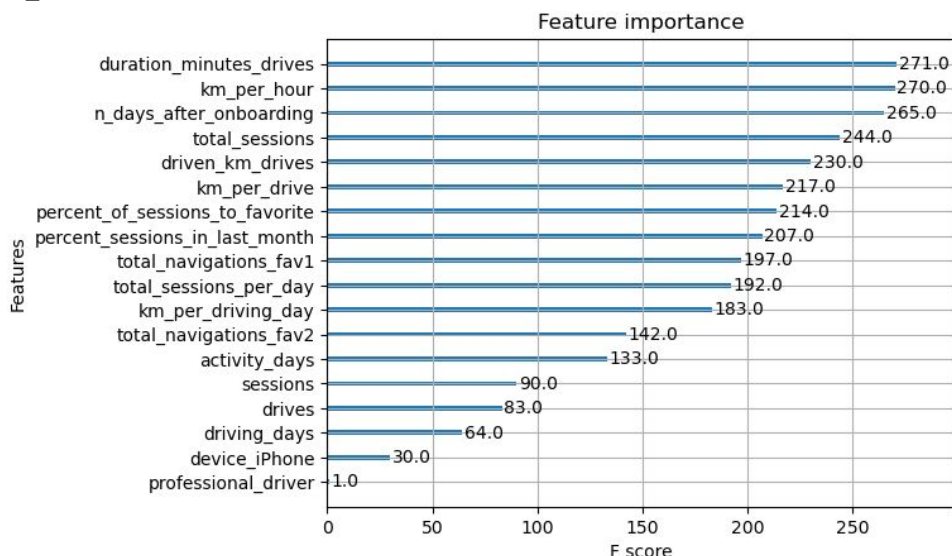
- **The model used six engineered features to rank the top 10** features, such as km_per_hour and percent_sessions_in_last_month.

- **The XGBoost model outperformed the random forest model** and improved the recall score by almost 100% compared to the logistic regression model from Milestone 5, while keeping similar accuracy and precision scores**.**

- **The tree-based models in this project milestone are superior to the logistic regression model because they have higher scores for all evaluation metrics and need less data preprocessing. However, they are harder to interpret.**