User Churn Project | ML Model Results

Prepared for: Waze Leadership Team

ISSUE / PROBLEM

The Waze data team is presently crafting a data analytics initiative focused on bolstering overall growth by mitigating monthly user churn in the Waze app. Within this project scope, churn is defined as users who have uninstalled or ceased app usage. The project's primary objective is to construct a machine learning (ML) model capable of predicting user churn.**This report offers details and key insights from ML modelling.**

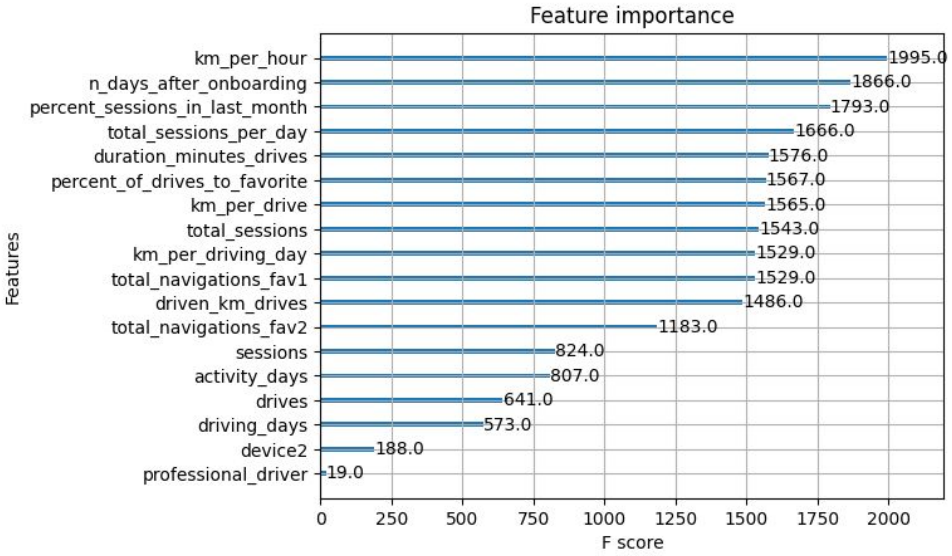
IMPACT

1. The ML models developed demonstrated a critical need for additional data in order to more accurately predict user churn.
2. This model results indicates that the existing data is inadequate for reliably forecasting churn. Incorporating drive-level details per user, like drive times and geographic locations, would enhance insights. Additionally, obtaining more detailed data on user interactions with the app, such as frequency of reporting road hazards, is advisable. Finally, tracking the monthly count of unique starting and ending locations inputted by each driver could provide valuable information.

RESPONSE

- **To obtain a model with the highest predictive power,two different Ensemble learning algorithms the random forest and XGBoost were used.**
- 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.**

KEY INSIGHTS



- **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 XGBoost was higher of the score from the logistic regression while still maintaining a similar accuracy and precision score.
- **The ensembles of tree-based models surpass the single logistic regression model due to superior performance across all evaluation metrics and reduced data preprocessing requirements. Nonetheless, comprehending the prediction mechanisms of tree-based models poses a greater challenge.**