User Churn Project | ML Model Results

ISSUE / PROBLEM

The Waze data team is currently developing a data analytics project aimed at increasing overall growth by preventing monthly user churn on the Waze app. For the purposes of this project, churn quantifies the number of users who have uninstalled the Waze app or stopped using the 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, which could impact the future development of the project, should further work be undertaken.**

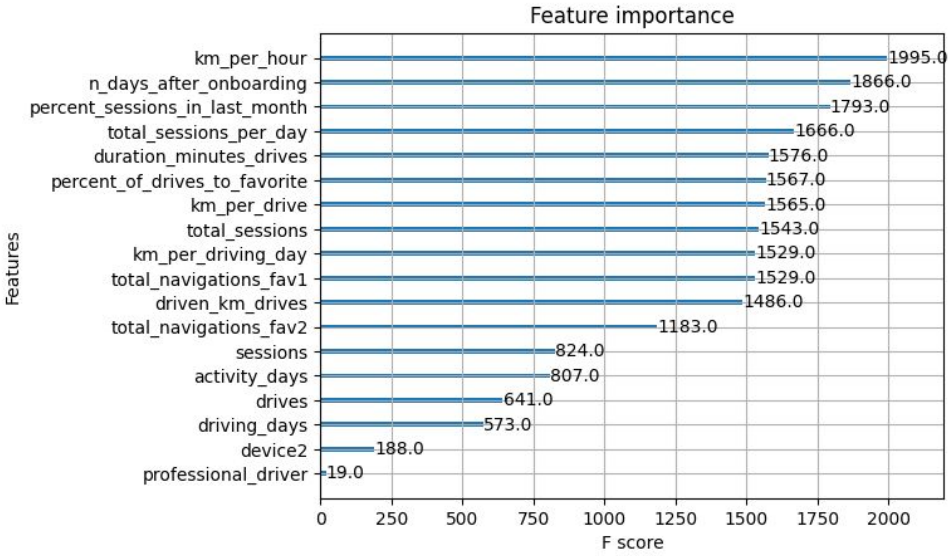
IMPACT

- ➔ **The ML models developed demonstrate a critical need for additional data in order to more accurately predict user churn.**
- ➔ **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. For example, how often do they 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.
- ➔ **Since engineered features are a proven valuable tool for improving the performance of ML models, a second iteration of the User Churn Project is recommended.**

RESPONSE

- **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.**

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 (17%) is nearly double the score from the previous logistic regression model built, while still maintaining a similar accuracy and precision score.
- **The ensembles of tree-based models in this project iteration 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.**