Waze: Predicting User Churn

Predictive Model

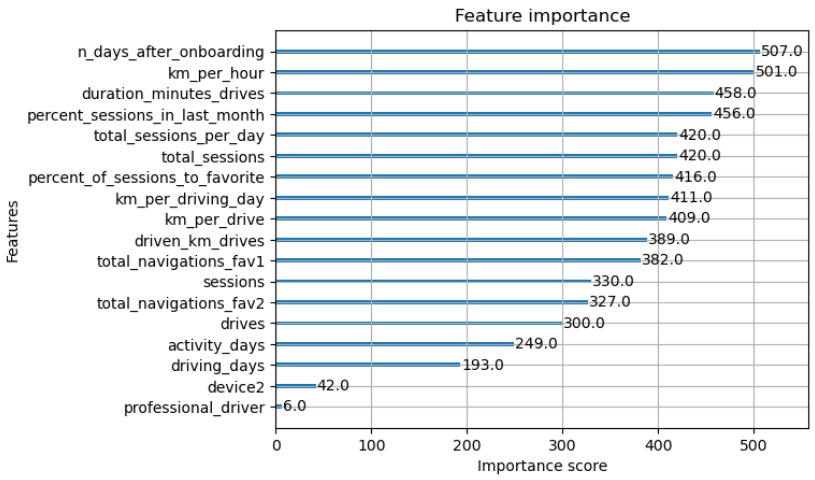
Project Overview

The objective of this project was to prepare the data for modelling, conduct feature engineering and to construct two ensemble tree-based models and evaluate their performance of classifying whether a user will churn or be retained. The ensemble models were a Random Forest Classifier and a XGBoost Classifier with the XGBoost being the champion model chosen after model selection on the validation set and further evaluated on the test set.

Key Insights

- The features: **km_per_driving_day**, **percent_sessions_in_last_month**, **professional_driver**, **total_sessions_per_day**, **km_per_hour**, **km_per_drive** and **percent_of_sessions_to_favorite**, were engineered to create a stronger predictive signal.
- The **outliers** were not imputed as the **tree-based models** are resilient to outliers.
- The **data** for the model was split into a **60/20/20 split** into training, validation and test sets.
- Both a **Random Forest Classifier** and a **XGBoost Classifier** was trained and evaluated with the **XGBoost Classifier** being chosen as the **champion model** on the validation set.
- The **champion XGBoost model** achieved a **Precision** of 37%, a **Recall** of 21%, **F1-Score** of 27% and an overall **Accuracy** of 80% on the test set.
- From the **Confusion Matrix** of the champion model on the test set, the model achieved **2166** True Negatives, **110** True Positives, **187** False Positives and **397** False Negatives.

Details



From the feature importance plot above we can observe that n_days_without_onboarding had the most influence on the model's predictions. Other features such as km_per_hour, duration_minutes_drives, percent_sessions_in_last_month and total_sessions_per_day were also important features in the model's predictions.

Next Steps

- To engineer new features to improve the model's performance to try to generate a better predictive signal. Some of the top predictors for this model's predictions were featured engineered, thereby proving their predictive improvement of performance.
- Decision threshold optimisation to improve the capture of users who will actually churn even if this means an increase in False Positives. To optimise the threshold for a higher Recall score, such as for a 50% Recall Score.
- Design and test retention interventions by collaborating with marketing and production teams to develop targeted retention strategies for high-risk users. Perform A/B testing to measure the effectiveness of these interventions.