Executive Summary - Waze Churn Prediction Project

Machine Learning Models Results

ISSUE / PROBLEM

The goal of the project is to develop a model to help predict user churn to improve user retention and grow Waze's business. The company will be able to optimize user retention strategy, enhance user experience, and make data driven decisions about product development.

This report refers to **Milestone 6** - Build a Machine Learning Model.

The purpose is to construct a Machine Learning model to predict user churn based on the variables present in the dataset and new engineered features.

Two tree-based models have been built and tested, a random forest and a XGBoost model.

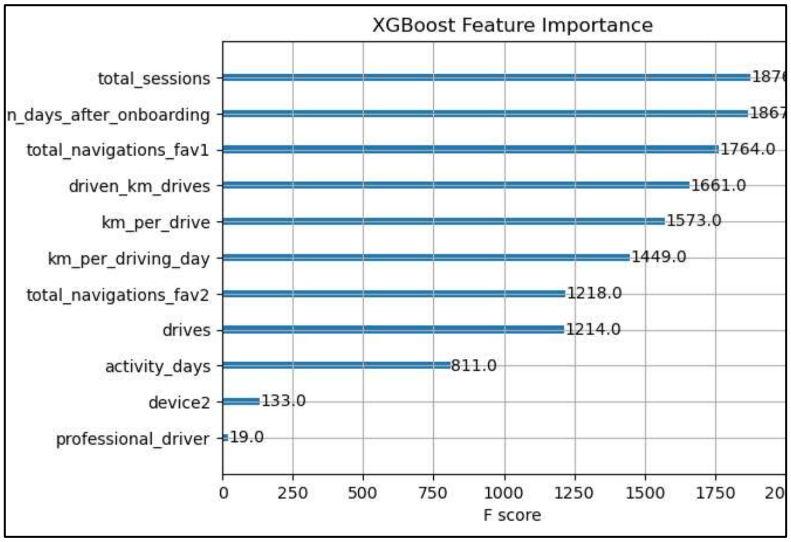
IMPACT

- ➔ The results of the analysis evidence that additional data is necessary to predict churn, given the poor predictive power of existing features
- ➔ The model could be marginally improved by adding new engineered features, but it appears that the features of the dataset are simply not good predictors.
- ➔ I suggest to evaluate the possibility of including more predictive features like telemetry data (e.g. new destinations counts, geographic locations, travel patterns, etc.)
- ➔ I would not recommend using the model for business decisions because the base prediction quality is mediocre (recall ~16%). Even optimizing recall to 50% (adjusting threshhold to 0.089) the number of false negatives is high compared to correct predictions of churned.

RESPONSE

- To get the best results two different models were built and tested: random forest and XGBoost.
- By splitting the data into 3 different datasets, the size of each dataset was reduced considerably. Starting from a dataset that was already quite small this impacts the quality of the training as it is done on a smaller set. However choosing a champion model on the validation set and testing it on the test data improves the estimation of future performance of the model on new data.
- The criteria for model selection is the "recall' metric which helps to better evaluate the false negatives (users predicted not to churn when in reality they churned).

KEY INSIGHTS



- The XGBoost model performed marginally better than the random forest model, and both performed better than the Logistic regression model.
- Engineered features accounted for six of the top 10 features (and three of the top five).
- The ensembles of tree-based models in this project milestone 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.