Title

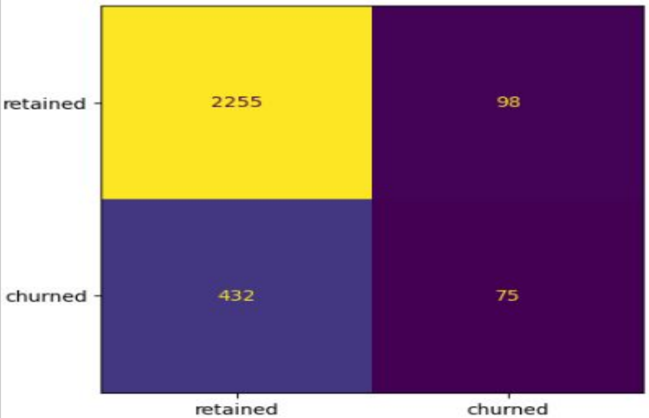
Waze’s Data Team Machine Learning Models Project

Project Overview

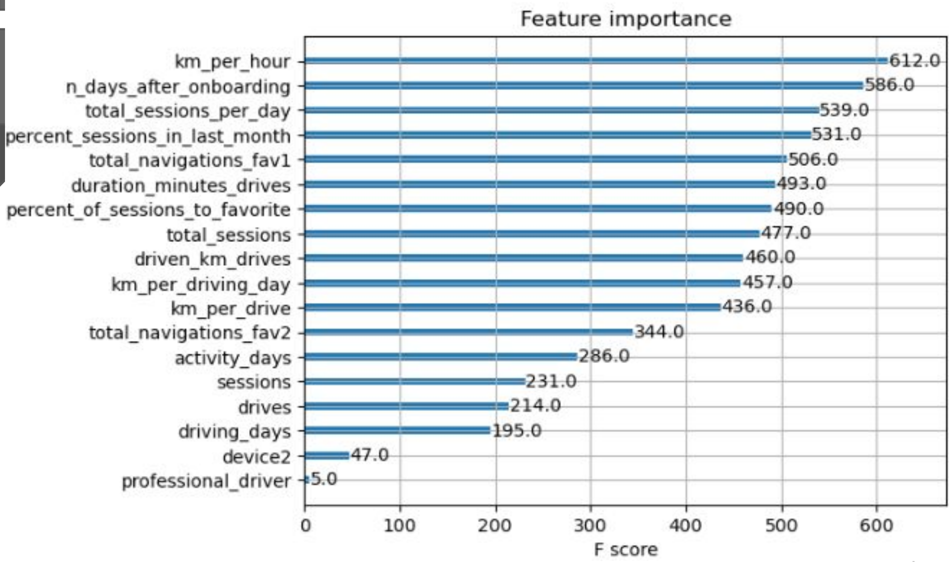
The purpose of this project is to develop machine learning models (Random Forest and XGBoost classification) to predict the churn user rate, and address features that have the most influence on whether or not users churn. Having the appropriate model will be tremendously beneficial to predicting the churn rate, and then come up with the measures to prevent, improve the retention rate, and grow Waze’s business.

Key Insights

- Both the **Random Forest** and **XGBoost** models address better results than the logistics regression. **XGBoost** is then chose as the final model to predict.
- However, this model wouldn’t improve much of the prediction results. Specifically, the main score metric is the recall score (which is also used in the logistic model to determine the model’s correctly identified churn rate) is about 0.1479, or 15%. If the model is used to deploy, it can only predict the 15% of people who are truly churned.
- Additionally, when using the test data for the model to predict, out of the 507 users who are truly churned, it correctly predicts 75 users, and incorrectly 432 users. This is the reason that the recall score is very low (refer to the plot below).
- Both **XGBoost** and **Random Forest** models surely can drive better results and better performance with little preprocessing work compared to the logistic one. However, they’re difficult to explain because of the complexity they bring.



Details



Top 10 features that have the most influence whether the users are churned. Half of them which are by creating new features (feature engineering) of existing ones (**km_per_hour**, **total_sessions_per_day**, **percent_sessions_in_last_month, km_per_driving_day**, and **percent_of_sessions_to_favorite**). It might be worth further examining these to better understand their effect on the target variable.

Next Steps

- Since this model doesn’t produce the best results, we won’t recommend deploying for business scenarios.
- The data team might further examine the effect of the model and do some more analysis.
- It might be worth accounting for collecting more data, and adding more features for the predictive signal of the target variable. We might once again build the model(s) to determine if it can drive to better performance and predictions.