ML models for churn prediction

Charly Moreno - Data Analyst (Waze's Data Analysis Team)

Project Overview

To improve the growth of the company through tackling users churn and improving the retention strategy, it is needed to fit and evaluate ML models (Random Forest -RF- and Extreme Gradient Boosting -XGBoost-) and provide feasibility of using them for this task.

Key Insights

The XGBoost model outperformed RF models for predicting churned users. However, its Recall and Precision scores doesn't qualify to be a production model.

This model predicts 53% of true positives (Recall score), and includes 17 features, being days after onboarding and activity days the most relevant ones.

It is suggested to generate a production quality model to set the retention strategy. This ideal model should have at least >75% score in Precision, Accuracy and F1, and >90% in Recall, since the aim is to predict churned users.

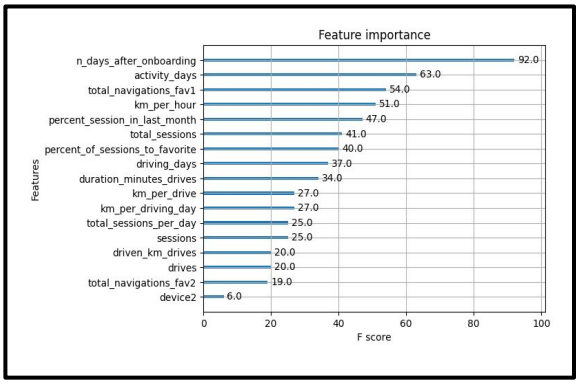
To do that, efforts should be done to **deal with the class imbalance** of the dataset (82% retained vs 18% churned users) and to **improve the quality of the data** (inconsistency in and lack of features).

Details

The data was splitted into training (60%), validation (20%) and test (20%) sets, to minimize bias. The first table shows the comparison of these three models.

	model	precision	recall	F1	accuracy
0	XG Boost	0.259693	0.627215	0.364454	0.612576
1	XG Boost (Val)	0.284838	0.641026	0.394417	0.651049
2	XG Boost (Test)	0.232699	0.530572	0.323512	0.606643

Evaluation metrics for XGBoost model (training, validation and test results)



Feature importance graph of the final XGBoost model (test results)

Six of the seventeen selected features of the final model (showed in the second table) were engineered.

Next Steps

To do a production level model:

- **Do a new data extraction**: Ideally dataset should include activity from different months using random sampling.
- **Increase feature engineering**: More predictor features when generating the models will improve the performance
- **Tackle class imbalance problem**: Use a sampling technique over the dataset.