Waze: Predicting User Churn

Binomial Logistic Regression Model

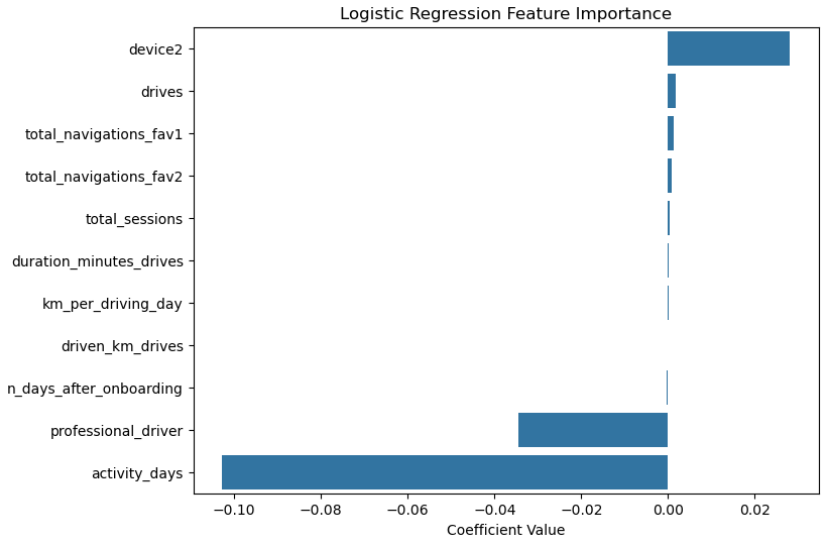
Project Overview

The objective of this project was to present knowledge of exploratory data analysis (EDA) and to build a logistic regression model. This included building a binomial logistic regression model, checking its assumptions and evaluating its classification performance for predicting user churn.

Key Insights

- The features **km_per_driving_day** and **professional_driver** were engineered. **professional_driver** was created from users who had 60 or more drives and drove on 15+ days in the last month.
- The **churn rate** for professional drivers was **7.6%** while the rate for non-professionals was **19.9%**.
- The **outliers** were changed to the **95th percentile** for their respective columns. This way they were capped to prevent a small number of outliers from disproportionately influencing the results.
- The model achieved an **Accuracy** of 83%, **Precision** of 57% and a **Recall** of 10% on the test set.
- From the **Confusion Matrix** on the test set, the model achieved **2315** True Negatives, **50** True Positives, **38** False Positives and **457** False Negatives.
- **activity_days**, **professional_driver** and **device2** had the most important influence on the model's predictions.

Details



From the feature importance plot above we can infer that activity_days was the most important influence to the model for its predictions. Other important features that have great influence on the models predictions are professional_driver and device2.

Next Steps

- Validate the logistic regression model on unseen data to reduce the risk of overfitting.
- Engineer new features to try to generate a better predictive signal as well as to scale the independent variables. This can include reconstructing the model with different combinations of the predictor variables as to reduce the noise from the unpredictive features.
- Test alternative models such as tree-based models to compare accuracy and interpretability.