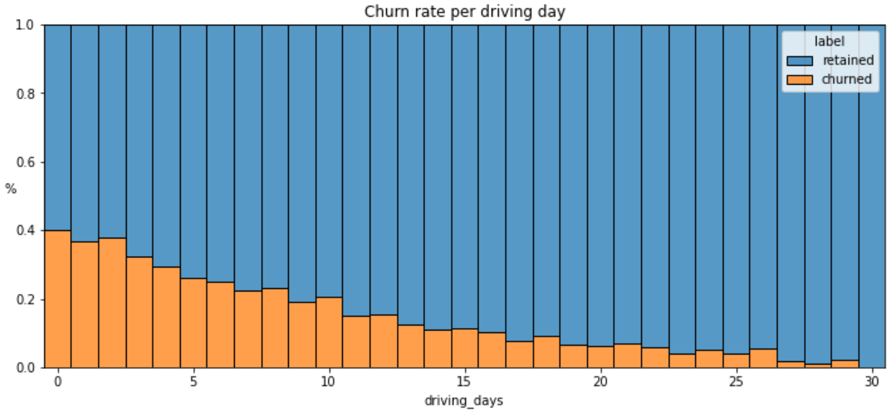
# Predicting User Churn for Waze

## Overview

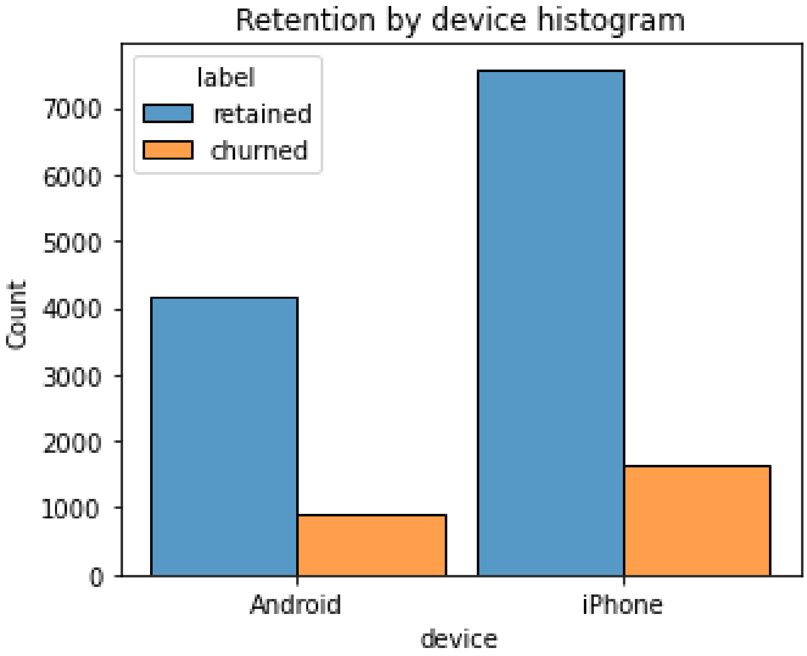
The Waze data team has been tasked with examining user churn on the Waze app, specifically those users who have uninstalled or stopped using the app. This can result in quantifiable actions that will increase overall growth. The data is composed of retained users (82%) and churned users (18%). The data has twelve variables consisting of objects, floats, and integers. 700 values are missing from the label column. Churned users drove 3 more drives, drove longer distances, and longer durations  than retained users. This is indicative of a particular user profile for the churned user which we recommend further focused analysis. Retained users utilized the app in double the amount of time than churned users. Median churned users drive twice the amount and 200 kilometers more (or 250% more) than the retained driver.

The data was predominantly right skewed (meaning all users had values in the lower end of that variable) or uniform (meaning that values were generally equal for that variable. Most of the data was not problematic but there were outliers in driven_km_drives, activity_days, and driving_days. Further questions to ask the Waze team might include why there was a surge in the number of long term users that suddenly started using the app within the last month. About 82% were retained users while 18% were churned users. Factors that correlated positively with user churn were the distance driven per day. The further a user drove each day, the higher the likelihood of that user churning. On the flip side, driving days had a negative correlation with churn as a user drove more days, they were less likely to churn. There was a uniform representation of new users in the data, shown in the histogram, n_days_after_onboarding.



## Business Understanding

The hypothesis test is insightful in that both groups of iPhone and Android users have a similar number of drives. Since the type of phone does not affect the number of drives, next steps to recommend would be to further investigate what other factors could contribute. It could be changes in marketing or UX interface.

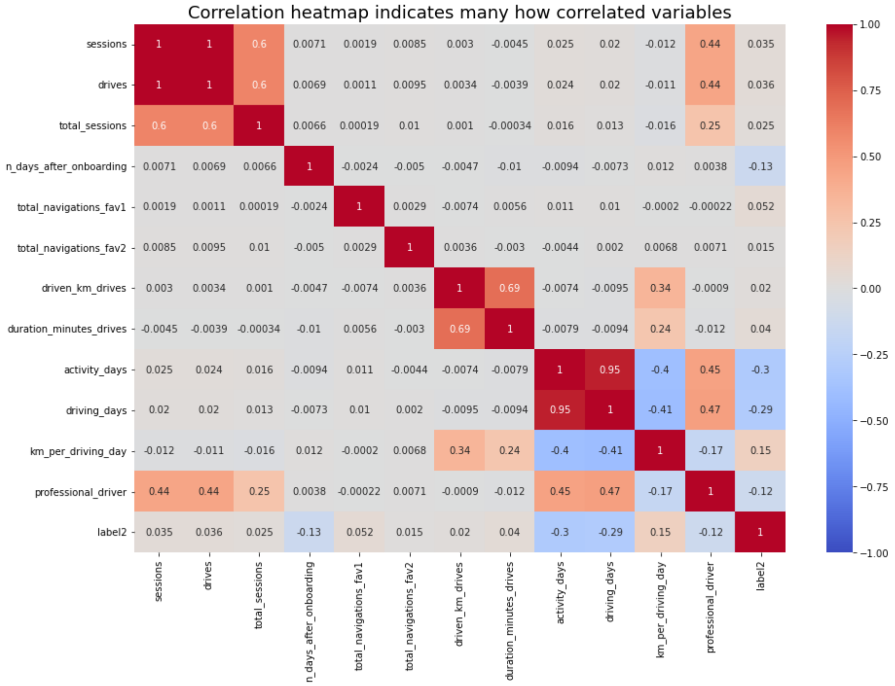


### Data Understanding

From the earlier EDA, we discovered that churn rate correlates with distance driven per driving day in the last month. It might be helpful to engineer a feature that captures this information.

The activity_days variable was the most influenced by the model's prediction because it had a negative correlation with user churn. This isn't a surprise as the variable is highly correlated with driving_days which was known to have a negative correlation with churn.

Yes, km_per_driving_day was expected to be a stronger predictor than they were, as shown in the previous EDA. However, in the model, it was shown to be the second least important variable.



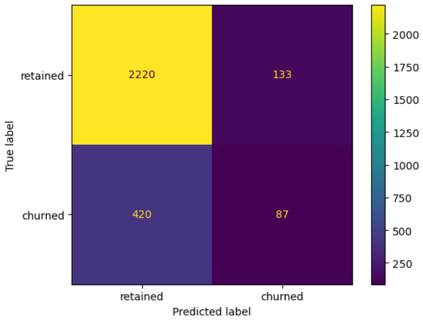
For classification tasks, tree-based models are better predictors than logistic regression models. This is due to that tree-based models require less data cleaning with fewer assumptions regarding the underlying distributions of their predictor variables, thus are easier to use.

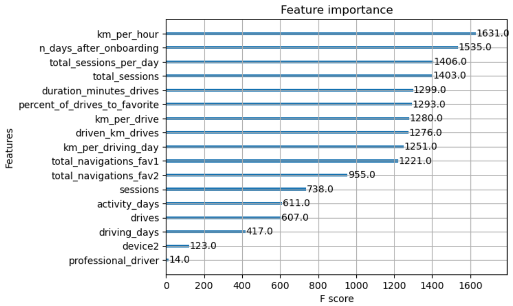
To improve this model, it would be beneficial to engineer new features to improve predictive signals. Particular to this case, the engineered features inhabit the top ten most predictive features. Another useful possibility would be to use different combinations of predictor variables to reduce the noise from predictive features.

Additional features that would help improve this model would be drive-specific information such as times, geographic locations, weather conditions. Secondly, more granular data on how users utilize the app such as how often they report road hazards. Thirdly, a monthly sum of unique starting and ending locations of each driver would be useful.

## Modeling and Evaluation

The prediction model determines to see if the customer will churn or will be retained. The ethical implications of the model if it makes errors is that Waze users will stop using the app, such as when the model predicts a false positive (the model says the Waze user will churn, but actually doesn't) is that it will increase annoyance or create a negative experience for the user. The benefits of a model that predicts users churning may be that it does just that. Thus, follow-up is recommended to ensure measures are applicable and the model benefits outweigh the problems.





## Conclusion

For classification tasks, tree-based models are better predictors than logistic regression models. This is due to that tree-based models require less data cleaning with fewer assumptions regarding the underlying distributions of their predictor variables, thus are easier to use. To improve this model, it would be beneficial to engineer new features to improve predictive signals. Particular to this case, the engineered features inhabit the top ten most predictive features. Another useful possibility would be to use different combinations of predictor variables to reduce the noise from predictive features. Additional features that would help improve this model would be drive-specific information such as times, geographic locations, weather conditions. Secondly, more granular data on how users utilize the app such as how often they report road hazards. Thirdly, a monthly sum of unique starting and ending locations of each driver would be useful.