Waze: Predicting User Churn

Data Preperation

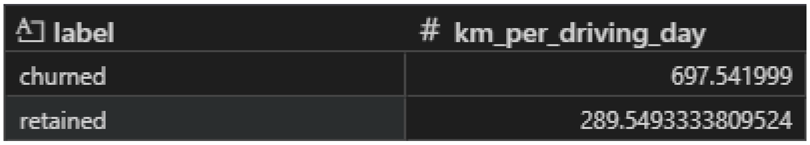
Project Overview

The objective is to prepare Waze user data to support a future churn prediction model. The analysis focused on understanding the data structure, identifying missing values, assessing variable distributions and comparing behavioural patterns between retained and churned users. This stage ensures that the dataset is suitable for EDA and predictive modelling.

Key Insights

- The dataset contains **14,999** users and **13** variables, with **700** missing values in the churn label only.
- The missing churn labels appear to be **missing at random**, with there being no meaningful differences in user behaviour or device type.
- Several of the usage variables show **extreme outliers** such as very high monthly KM driven, thereby making the median more representative than the mean.
- Churned users exhibit **more intense but less frequent driving behaviour**, while the retained users **show more regular engagement across more days**.
- The device type (Android vs iPhone) shows no appreciable difference in churn rate.

Details



label	# km_per_driving_day
churned	697.541999
retained	289.5493333809524

- **Missing Data:** Only the `label` column contains the missing values while all the other variables are complete. Device distribution amoung missing labels closely matches the overall dataset, thereby indicating no systemic bias.
- **Distribution Shape:** Driving related variables are highly right-skewed with extreme maximum values, suggesting the presence of high-intensity or professional drivers.
- **Churn vs Retention Patterns:** The churned users drive further per driving day and complete more drives per day, but use the app on fewer days. In comparision, the retained users interact with the app on more days, suggesting steadier engagement.
- **Device Analysis:** The Android/iPhone split is consistent across retained and churned users.

Next Steps

- Conduct deeper **EDA** to explore skewness, outliers and behavioural segments.
- Consider **user segmentation** such as commuters vs high intensity drivers to better understand the drivers of user churn.
- **Engineer normalised features** such as KM per driving day, for modelling.
- Evaluate whether extreme users should be **modelled separately or capped** to prevent distortion.
- Proceed to **feature selection** and **baseline churn modelling**.