User Churn Project: A Case study of Waze

# Introduction and Background

We downloaded a dataset on Waze app users' activities, a fictional transportation company from Kaggle. The dataset contained the following columns,

* ID (unique ID for each customer)
* Label (retained or churned)
* Sessions (number of times the user logged in to the app)
* Drives (number of times the user drove the vehicle)
* total\_sessions (total number of times the user logged in to the app since its installation)
* n\_days\_after\_onboarding (number of days since the user first logged in to the app)
* total\_navigations\_fav1 (number of sessions/visits to favorite place 1)
* total\_navigations\_fav2 (number of sessions/visits to favorite place 2)
* driven\_km\_drives (kilometers driven so far)
* duration\_minutes\_drives (minutes the vehicle was drove)

**Goal: This model aims to find factors that drive user churn and** to predict whether a Waze user is retained or churned.

This analysis will help Waze's initiative to enhance growth. High retention rates usually indicate satisfied users who consistently use the app over time. By developing a churn prediction model, we aim to prevent user churn, improve retention, and support Waze's business growth.

# Key Questions

1. What is the current churn rate?
2. How many users are actively using the app?
3. How much time does an average user spend on the app?
4. Who are the most at-risk users?
5. How can we prevent user churn?
6. How will project success be measured?

# Detailed Analysis Approach

Before developing the predictive model, we'll undergo a thorough analyze stage to test several hypotheses, gain insights, and uncover potential relationships within the data. This stage is crucial for laying the groundwork for accurate and reliable predictions.

1. **Identifying Data Anomalies:**We will begin by examining the dataset for any anomalies, such as patterns, missing values, and outliers. These anomalies can significantly impact the analysis and the performance of the predictive model. Identifying them early helps us understand the data's integrity and make necessary adjustments.
2. **Data Cleaning:**The next step involves cleaning the dataset to ensure it is ready for analysis. This process includes removing duplicates, addressing missing values, and eliminating outliers that could skew the results. Once the data is cleaned and properly formatted, we will proceed with exploratory data analysis (EDA) to gain a deeper understanding of the dataset's structure and key characteristics.
3. **Basic Statistical Analysis:**We will conduct fundamental statistical analyses, such as correlation and regression analysis, to explore relationships and causality between variables. This step helps in identifying which factors are most influential in predicting user behavior and potential churn.
4. **Machine Learning Model Development:**At this stage, we will implement machine learning models using two or three different algorithms to determine the most effective approach. By comparing the performance of these models, we can identify the best strategy to address the issue of user churn. Model selection will be based on accuracy, precision, recall, and other relevant metrics.
5. **Final Recommendations and Report:**After completing the analysis and model development, we will provide actionable recommendations to the company. These suggestions will be based on our findings and will aim to improve user retention and reduce churn. Additionally, a detailed report will be written, summarizing the entire project, including the methodologies used, results obtained, and recommendations for future actions. This report will serve as a comprehensive resource for the company to implement the suggested strategies.

# Project Implementation

1. Imported the necessary libraries, loaded the data, and examined the top and bottom rows of the dataset.
2. Checked the datatypes and identified 700 null values in the ‘label’ column.
3. Segregated the dataset into null and non-null values and assessed if there were significant differences or proportional distribution of device types in the null dataset. Found no significant differences, justifying the removal of rows with null values.
4. Analyzed the percentage of retained and churned users in the overall dataset.
5. Determined that outliers were present, making the median a better measure than the mean. Calculated median values for each ‘label’ category, finding that churned users had more drives, greater kilometers driven, and longer drive durations on average compared to retained users.
6. Observed that churned users drove more distance per day, spent more time on the road, and had a higher number of sessions per day compared to retained users.
7. Compared iPhone and Android users in both churned and retained groups, finding no significant differences.
8. Conducted exploratory data analysis (EDA) to gain a deeper understanding of the dataset.
9. Added additional visualization libraries, such as Seaborn and Matplotlib.
10. Created functions for histograms and boxplots, and used a loop to generate graphs for all variables.
11. Detected and removed outliers by creating a function to establish thresholds.
12. Used logistic regression as the initial model, focusing on the primary binary variable ‘label’.
13. Added new variables (kilometers per day and driver type) and encoded labels and devices to binary values.
14. Checked for multicollinearity and found a high correlation between activity days and driving days.
15. Applied logistic regression, including data segregation (X and y), training/testing split, and model evaluation.
16. Conducted additional feature engineering before proceeding with XGBoost, adding features like the percentage of sessions in the last month etc.
17. Re-split the data into training and testing sets, trained the XGBoost model using optimal parameters.
18. After building and testing the machine learning models, the next step was to share the findings and assess their value.

# Results

1. Discovered 700 null values in the ‘label’ column, which were safely removed after confirming no significant differences between datasets.
2. Found that churned users had more drives, drove more kilometers, and spent more time on the road on average compared to retained users.
3. Identified that outliers could distort results, making the median a more accurate measure than the mean.
4. Observed no significant differences between iPhone and Android users in both churned and retained groups.
5. The initial logistic regression model had an accuracy of 0.83, a precision of 0.49, a recall of 0.06, and an F1 score of 0.09.
6. XGBoost model showed an improvement in recall (0.13) and F1 score (0.19) but a decrease in precision (0.36) and accuracy (0.81).
7. Determined that while the model had potential value for exploratory purposes, it was not suitable for guiding critical business decisions due to its poor recall score.
8. Highlighted the need for additional data, such as drive-level information and user interaction details, to enhance model performance.
9. Recognized that logistic regression models provide greater interpretability, while tree-based models like XGBoost offer higher predictive power with fewer assumptions.

# Conclusion and Recommendation for Waze Management

Based on the analysis and results, the current models developed for churn prediction offer limited practical value for guiding critical business decisions at Waze. The low recall score, particularly in the logistic regression model, indicates that the model is not effective at correctly identifying users who are likely to churn. While the XGBoost model showed some improvement in recall and F1 scores, the overall predictive power remains insufficient for reliable churn prediction, especially when precision is also a concern.

# Recommendations for future researchers

1. **Model Improvement and Data Enhancement:**
   * **Feature Engineering:** Further refine the model by engineering new features that may have a stronger correlation with user churn. This could involve exploring user behavior metrics like drive times, geographic locations, and more detailed interaction data within the app.
   * **Additional Data Collection:** To improve model performance, consider collecting more granular data, such as drive-level information and user interactions. This additional data can offer deeper insights and enhance the model's ability to predict churn more accurately.
2. **Consideration of Model Types:**
   * **Logistic Regression:** Continue to use logistic regression for its interpretability, which can be valuable in understanding which features most influence user churn.
   * **Tree-Based Models:** When accuracy is a priority, tree-based models like random forests or XGBoost should be considered, as they typically require less data cleaning and handle complex relationships between variables more effectively.
3. **Future Development:**
   * Invest in improving the existing models by leveraging domain knowledge and advanced feature engineering.
   * Explore alternative machine learning models that might better capture the complexity of user behavior and provide more accurate churn predictions.