User Churn Project – Data Summary

Waze leadership wants to build a machine learning model to predict user churn. The model is based on data collected from users of the Waze app. This project aims to increase user retention and support overall growth of Waze.

This document includes initial key insights from the second milestone, revealed details with analysis, and recommendation about the next steps.

Key Insights

- **The data contains missing values only in 'label' column. There are 700 values missing with no indication that the omissions are non-random.**
- **The percentage of retained users is 82 percent while the percentage of churned customers is 18 percent approximately.**
- **Churned users has driven 608 km per day. It is a significantly higher value compared to retained users who has driven 245 km per day.**
- **The iPhone users is represented by 65 percent of the data while Android users had 35 percent approximately.**
- **The median churned user drove 200 more kilometers and 2.5 more hours during the last month than the median retained user.**

Details

Milestone 2 - Compile summary information

🎯 **Target Goal:** Inspect user data to learn important relationships between variables.

🎯 **Methods:**

- Built a dataframe
  - Each row represents a single observation, and each column represents a single variable
- Collected preliminary statistics
- Analyzed user behavior

🎯 **Impact:** Our team determined important relationships between variables that will guide further analysis of user data.

Next Steps

➔ **Our team recommends gathering more data on the super-drivers**. It's possible that the reason they're driving so much is also the reason why the Waze app does not meet their specific set of needs, which may differ from the typical driver.

➔ **The immediate next step is to conduct thorough EDA and develop data visualizations** to illustrate the narrative behind the data and guide future project decisions.

User Churn Project | Exploratory Data Analysis

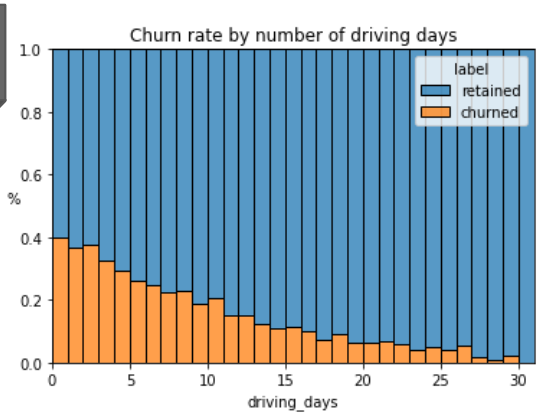
Project Overview

The Waze team is currently developing a data analytics project support overall growth by decreasing monthly user churn on the Waze app. This report contains insights from Exploratory Data Analysis process.

Key Insights

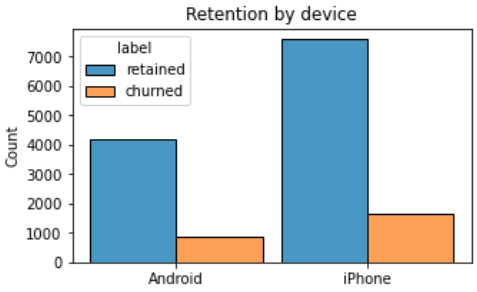
- **Less then %18 of user churned** while more than %82 of user retained.
- **Distance driven per driving day and user churn rate have a positive correlation**. Users drove more each driving day are more likely to churn.
- **Number of driving days has a negative correlation with user churn.** Users who drove more days are less likely to churn.
- Almost every variable is **uniformly distributed or extremely right-skewed.**
- For **right-skewed distri**bution this means that **half of the values is positioned in the left side on the range.**
- For **uniformly distribution** this mean this means that **values are equally distributed within the range.**

Details



The people who used Waze less during the month have higher churn rate.

Retention rate stayed consistent between device types.



Next Steps

- ➔ **Inspect the problematic values in ‘number_of_sessions’, ‘driving_days’, ‘activity_days’**
- ➔ **Explore user profiles to gain insights on the reason for the long-distance driver’s churn rate.**
- ➔ **Run deeper analysis for the impacts of variables on user churn.**

User Churn Project | Two-Sample Hypothesis Test Results

Overview

The Waze data team is currently developing a data analytics project aimed at increasing overall growth by preventing monthly user churn on the Waze app.
As part of the effort to improve retention, Waze wants to learn more about users' behavior.

Objective

Determine whether there is statistically significant difference in average amount of rides between iPhone and Android users by conducting a hypothesis test using a two-sample t-test.

Results

Based on the initial calculation, we observe that **iPhone users slightly have larger mean number of rides.**

	device	mean amount of rides
0	Android	66.231838
1	iPhone	67.859078

The p-value is greater than the significance level, therefore we fail to reject the null hypothesis. **There is no statistically significant difference in average amount of rides between iPhone and Android users.**

The average amount of rides is similar between iPhone and Android users.

Next Steps

Potential next step is to explore what other factors influence the variation in the number of drives and run additional hypothesis tests to learn more about user behavior. Further, temporary changes in marketing or user interface for the Waze app may provide more data to investigate churn.*

User Churn Project | Regression Modeling Results

Overview

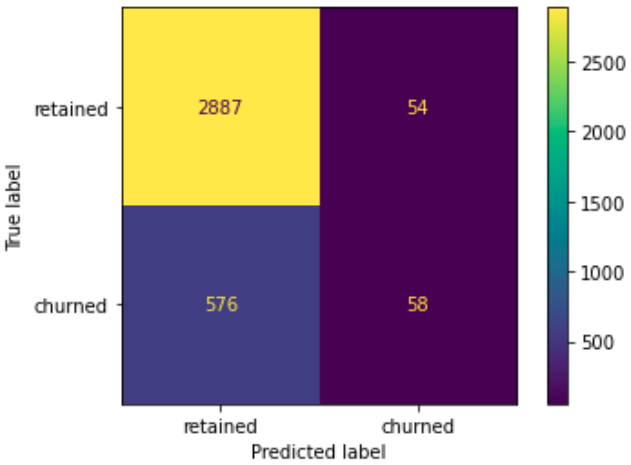
The Waze data team wants to build a binomial logistic regression model to predict user churn because it is well=suited for modeling such binary outcomes and provides clear and interpretable results. Logistic regression can also handle imbalanced datasets well and still provide meaningful predictions to inform business decisions.

Objective

The objective is applying user data to build a binomial logistic regression model.

- Create features **km_per_driving_day** and **professional_driver**
- Encode categorical variables **label** and **device**
- Assess features for multicollinearity.
- Build the regression model.

Results



The efficacy of a binomial logistic regression model is determined by accuracy, precision, and recall scores; in particular, **recall is essential to this model as it shows the number of churned users.**

The precision score was 52% which is mediocre however, the recall score was 9% which indicates that the model fails to capture churned users. This means that model makes a lot of false negative predictions and fails to identify churned users.

'activity_days' was the most important feature of the dataset, and it had a negative correlation between user churn.
In previous EDA, user churn rate increased as the values in km_per_driving_day increased. **In the model, distance driven per day was the second-least-important variable.**

Next Steps

While this model may not be suitable for making critical business decisions, it provides valuable insights. It highlights the importance of acquiring additional data (features) that show correlation with user churn. Furthermore, it suggests a potential need to refine the user profile that Waze aims to engage with, aligning with their goal of enhancing overall growth by mitigating monthly user churn on the app.

User Churn Project | ML Model Results

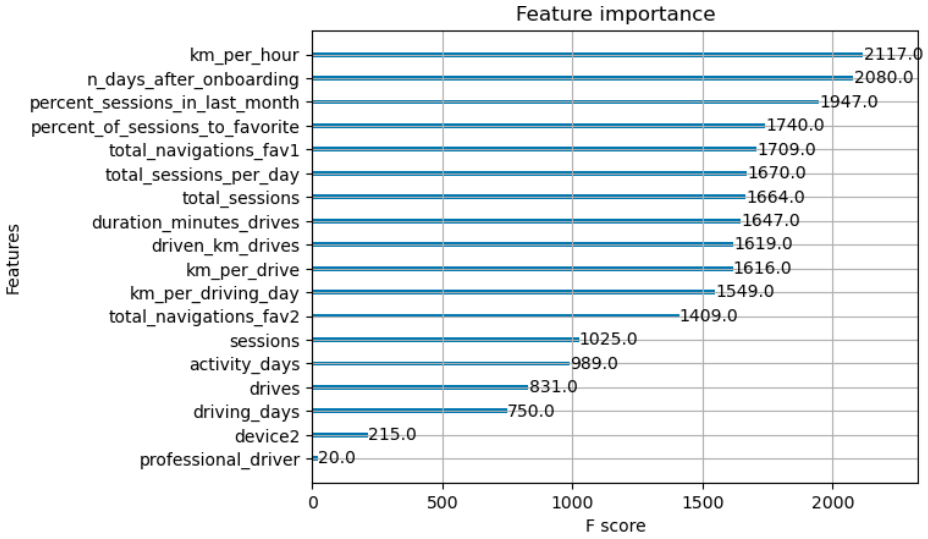
ISSUE / PROBLEM

The Waze data team is working on a project to reduce monthly user churn on the app. The goal is to develop a machine learning model predicting user churn. This report summarizes key insights from Milestone 6, which could impact future project development.

RESPONSE

- To enhance predictive power, the Waze data team created and compared two models – **random forest and XGBoost. The dataset was strategically divided into training, validation, and test sets.** Although the three-way split reduces the amount of data for model training compared to a two-way split, it allows for effective model selection on a separate validation set. This approach ensures that the champion model, identified through validation, is independently tested on the dedicated test set, providing a more robust estimate of future performance compared to a two-way split.

KEY INSIGHTS



- In the third milestone, engineered features, including km_per_hour, percent_sessions_in_last_month, total_sessions_per_day, percent_of_drives_to_favorite, km_per_drive, and km_per_driving_day, were instrumental, constituting six of the top 10 features.

- The XGBoost model outperformed the random forest model, showcasing a notable improvement in the recall score to 17%, nearly double that of the logistic regression model in Milestone 5. Despite maintaining similar levels of accuracy and precision, these ensemble models proved more valuable, requiring less data preprocessing.

- However, the complexity of tree-based models makes their predictions harder to interpret compared to singular logistic regression models.

IMPACT

- ➔ **The machine learning models developed in Milestone 6 have brought to light a significant requirement for additional data to markedly improve the accuracy of predicting user churn within the Waze app.**
- ➔ **Notably, the current dataset exhibits limitations that hinder consistent churn prediction, emphasizing the importance of incorporating drive-level information, more detailed insights into user interactions (such as reporting road hazards), and obtaining monthly counts of unique starting and ending locations for each driver.**
- ➔ **Given the proven effectiveness of engineered features in elevating the performance of machine learning models, the Waze team suggests undertaking a second iteration of the User Churn Project. This iteration aims to integrate the recommended additional data elements and further refine the modeling approach.**