User Churn Project | Preliminary Data Summary

Prepared for: Waze Leadership Team

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. For the purposes of this project, churn quantifies the number of users who have uninstalled the Waze app or stopped using the app.

This report offers a preliminary data summary, information on the project status and key insights of Milestone 2, which impact the future development of the overall project.

PROJECT STATUS

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.

KEY INSIGHTS

- This dataset contains **82% retained users** and **18% churned users**.
- The dataset contains 12 unique variables with types including objects, floats, and integers; the label column is missing 700 values with no indication that the omissions are non-random.
- Churned users averaged ~3 more drives in the last month than retained users.
- Retained users used the app on over twice as many days as churned users in the last month.
- The median churned user drove ~200 more kilometers and 2.5 more hours during the last month than the median retained user.
- Churned users had more drives in fewer days, and their trips were farther and longer in duration. Perhaps this is suggestive of a user profile; our team will have to continue exploring!
- The median user who churned drove 698 kilometers each day they drove last month, which is about 240% the per-drive-day distance of retained users.
- Regardless of user churn, the users represented in this data drive a lot! It is probably safe to assume that this data does not represent typical drivers at large.

User Churn Project | Exploratory Data Analysis

Prepared for: Waze Leadership Team

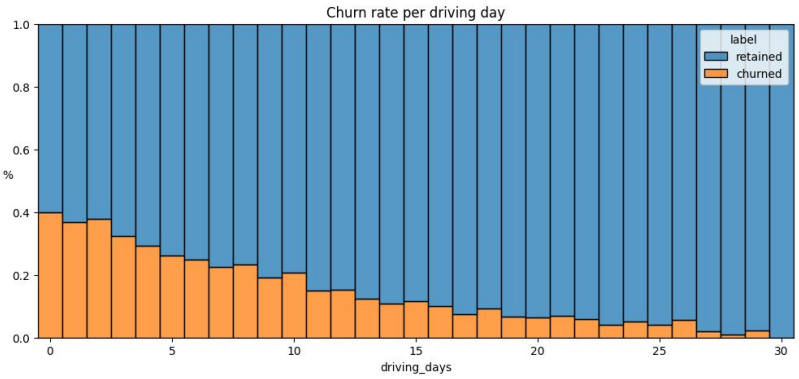
Project Overview

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. Thorough exploratory data analysis (EDA) enables Waze to make better decisions about how to proactively target users likely to churn, thereby improving retention and overall customer satisfaction. **This report offers details and key insights from Milestone 3, which impact the future development of the overall project.**

Key Insights

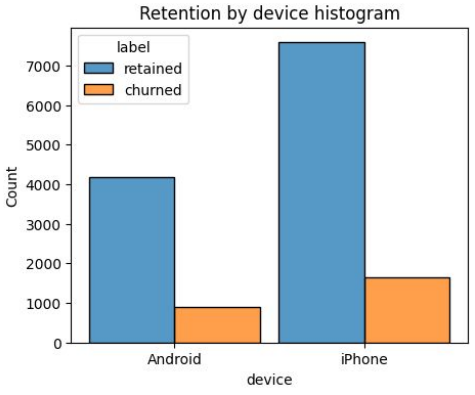
- **The more times users used the app, the less likely they were to churn.** While 40% of the users who didn't use the app at all last month churned, nobody who used the app 30 days churned.
- **Distance driven per driving day had a positive correlation with user churn.** The farther a user drove on each driving day, the more likely they were to churn.
- **Number of driving days had a negative correlation with churn.** Users who drove more days of the last month were less likely to churn.
- **Users of all tenures from brand new to ~10 years were relatively evenly represented in the data.**
- **Nearly all the variables were either very right-skewed or uniformly distributed.**
  - For the right-skewed distributions, this means that most users had values in the lower end of the range for that variable.
  - For the uniform distributions, this means that users were generally equally likely to have values anywhere within the range for that variable.
- **Several variables had highly improbable or perhaps even impossible outlying values**, such as: driven_km_drives, activity_days and driving_days.

Details



The churn rate is highest for people who didn't use Waze much during the last month.

The proportion of churned users to retained users is consistent between device types.



Next Steps

➔ **Investigate the erroneous or problematic discrepancies between number of sessions, driving_days, and activity_days.**

➔ **Continue to explore user profiles with the greater Waze team; this may glean insights on the reason for the long distance drivers' churn rate.**

➔ **Plan to run deeper statistical analyses on the variables in the data to determine their impact on user churn.**

User Churn Project | Two-Sample Hypothesis Test Results

Prepared for: Waze Leadership Team

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. **This report offers information on the project status and results of Milestone 4, which impact the future development of the overall project.**

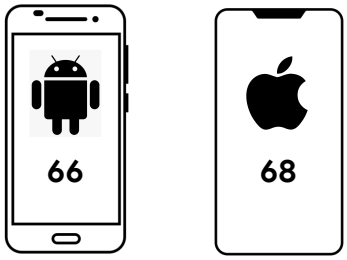
Objective

🎯 **Target Goal:** Develop a two-sample hypothesis test to analyze and determine whether there is a statistically significant difference between mean number of rides and device type – Android vs. iPhone.

🎯 **Impact:** Statistical tests, such as the one conducted for Milestone 4, enable the Waze data team to make inferences about the populations from which the data was drawn and help them learn more about their user base.

Results

Average Number of Drives



Note: The mean number of drives shown here – 66 for Android and 68 for iPhone – have been rounded up.

- Based on the calculations, drivers who use an iPhone to interact with the application have a higher number of drives on average.
- **The t-test results concluded there is not a statistically significant difference in mean number of rides between iPhone users and Android users.**

Next Steps

➔ **Due to the results rendered from this specific hypothesis test, the Waze data team recommends running additional t-tests on other variables to learn more about user behavior.**

➔ **Additionally, since the user experience is the same, temporary changes in marketing or user interface may be impactful rendering more data to investigate user churn behavior.**

User Churn Project | Regression Modeling Results

Prepared for: Waze Leadership Team

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. For the purposes of this project, churn quantifies the number of users who have uninstalled the Waze app or stopped using the app. Binomial logistic regression models typically offer flexibility and predictive power, which can be used to inform larger business decisions. Our team sought to build one from the data provided for this project. **This report offers details and key insights from Milestone 5, which impact the future development of the overall project.**

PROJECT STATUS

Milestone 5 - Regression Modeling

🎯 **Target Goal:** Apply user data to build and analyze a binomial logistic regression model.

🎯 **Methods:**

- Created features of interest to the stakeholders and business scenario
- Assessed features for multicollinearity
- Built the regression model
- Evaluated model performance

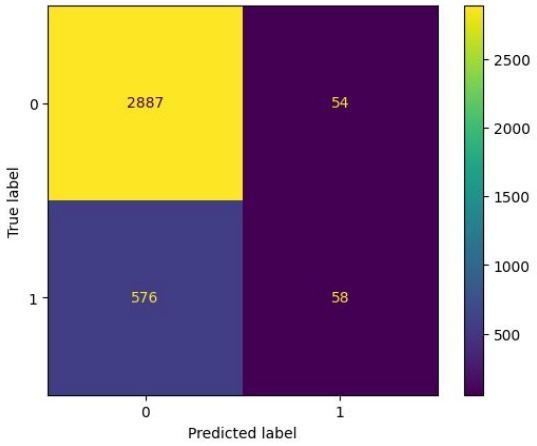
🎯 **Impact:** With enough data, binomial logistic regression model results can reveal important variable relationships and predict binary outcomes, which can inform decisions for marketing and product development, for example.

NEXT STEPS

➔ **Due to the model results, our team recommends using the key insights from this project milestone to guide further exploration.**

➔ **This model should not be used to make significant business decisions; however, it has valuable insights insofar as it demonstrated a great need for additional data (features) that correlates with user churn, and also a possible need to better define the user profile Waze seeks to target in their aim to increase overall growth by preventing monthly user churn on the app.**

KEY INSIGHTS



Note: 1 = churned and 0 = retained

- 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 model has mediocre precision (53% of its positive predictions are correct) but very low recall, with only 9% of churned users identified.** This means the model makes a lot of false negative predictions and fails to capture users who will churn.
- **Activity_days was by far the most important feature in the model.** It had a negative correlation with 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.**

User Churn Project | ML Model Results

Prepared for: Waze Leadership Team

ISSUE / PROBLEM

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. For the purposes of this project, churn quantifies the number of users who have uninstalled the Waze app or stopped using the app. The ultimate goal for this project is to develop a machine learning (ML) model that predicts user churn. **This report offers details and key insights from Milestone 6, which could impact the future development of the project, should further work be undertaken.**

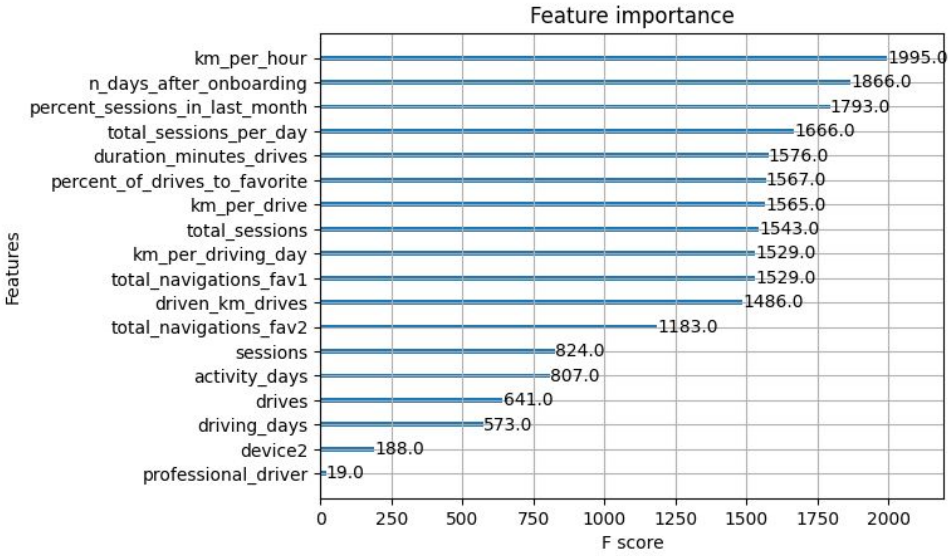
IMPACT

- ➔ **The ML models developed for Milestone 6 demonstrate a critical need for additional data in order to more accurately predict user churn.**
- ➔ **This modeling effort confirms that the current data is insufficient to consistently predict churn.** It would be helpful to have drive-level information for each user (such as drive times, geographic locations, etc.). It would probably also be helpful to have more granular data to know how users interact with the app. For example, how often do they report or confirm road hazard alerts? Finally, it could be helpful to know the monthly count of unique starting and ending locations each driver inputs.
- ➔ **Since engineered features are a proven valuable tool for improving the performance of ML models, the Waze team recommends a second iteration of the User Churn Project.**

RESPONSE

- **To obtain a model with the highest predictive power, the Waze data team developed two different models to cross-compare results: random forest and XGBoost.**
- To prepare for this work, the data was split into training, validation, and test sets. Splitting the data three ways means that there is less data available to train the model than splitting just two ways. However, **performing model selection on a separate validation set enables testing of the champion model by itself on the test set, which gives a better estimate of future performance than splitting the data two ways and selecting a champion model by performance on the test data.**

KEY INSIGHTS



- **Engineered features accounted for six of the top 10 features:** km_per_hour, percent_sessions_in_last_month, total_sessions_per_day, percent_of_drives_to_favorite, km_per_drive, km_per_driving_day.
- **The XGBoost model fit the data better than the random forest model.** Additionally, it's important to call out that the recall score (17%) is nearly double the score from the previous logistic regression model built in Milestone 5, while still maintaining a similar accuracy and precision score.
- **The ensembles of tree-based models in this project milestone are more valuable than a singular logistic regression model because they achieve higher scores across all evaluation metrics and require less preprocessing of the data. However, it is more difficult to understand how they make their predictions.**