README: Waze User Churn Prediction Project

Overview

The goal of this project was to develop a machine learning model to predict monthly user churn on the Waze navigation app. Churn is defined as users who have either uninstalled the app or stopped using it during a given month. The project progressed through exploratory data analysis, hypothesis testing, logistic regression, and tree-based machine learning models. The final XGBoost model achieved 17% recall - nearly double the logistic regression baseline - while maintaining comparable accuracy and precision. Key predictive features were predominantly engineered variables, including `km_per_hour`, `percent_sessions_in_last_month`, and `total_sessions_per_day`. The analysis revealed that current data is insufficient to reliably predict churn, and recommended collecting more granular user interaction data for future iterations.

Business Understanding

Waze's free navigation app serves millions of drivers worldwide, with growth dependent on retaining active users. High retention rates signal satisfied users who consistently engage with the app over time. Understanding and predicting churn enables Waze to proactively engage at-risk users through targeted interventions, ultimately improving retention and supporting business growth.

This project aimed to answer three critical questions:

- **Who** are the users most likely to churn?
- **Why** do users churn?
- **When** do users churn?

Accurate churn prediction allows Waze to identify at-risk user segments before they leave, enabling data-driven decisions about product development, marketing strategies, and user experience improvements.

Data Understanding

The dataset (`waze_dataset.csv`) contains synthetic data created in partnership with Waze, comprising **14,999 unique users** and **13 features** capturing user behavior and engagement metrics.

Data Dictionary

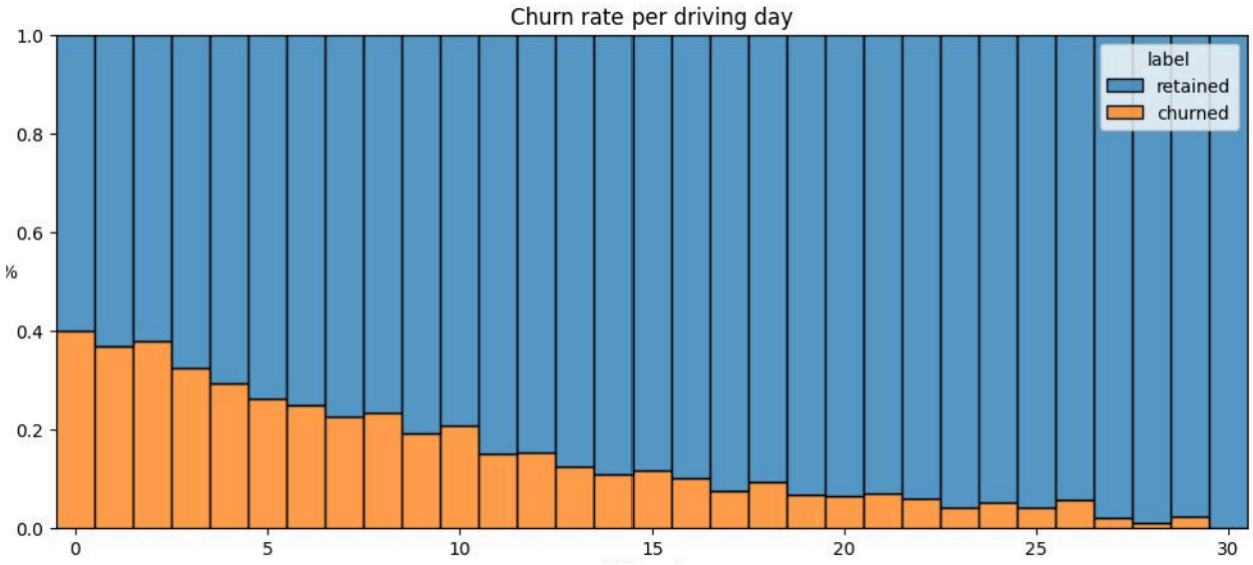
Column	Type	Description
ID	int	Sequential numbered index
label	obj	Binary target variable ("retained" vs "churned") indicating if a user churned during the month
sessions	int	Number of times the user opened the app during the month
drives	int	Number of occurrences of driving at least 1 km during the month
device	obj	Device type used to start a session (Android/iPhone)
total_sessions	float	Model estimate of total sessions since user onboarded
n_days_after_onboarding	int	Number of days since the user signed up for the app
total_navigations_fav1	int	Total navigations to the user's favorite place 1 since onboarding
total_navigations_fav2	int	Total navigations to the user's favorite place 2 since onboarding
driven_km_drives	float	Total kilometers driven during the month
duration_minutes_drives	float	Total duration driven in minutes during the month
activity_days	int	Number of days the user opened the app during the month
driving_days	int	Number of days the user drove at least 1 km during the month

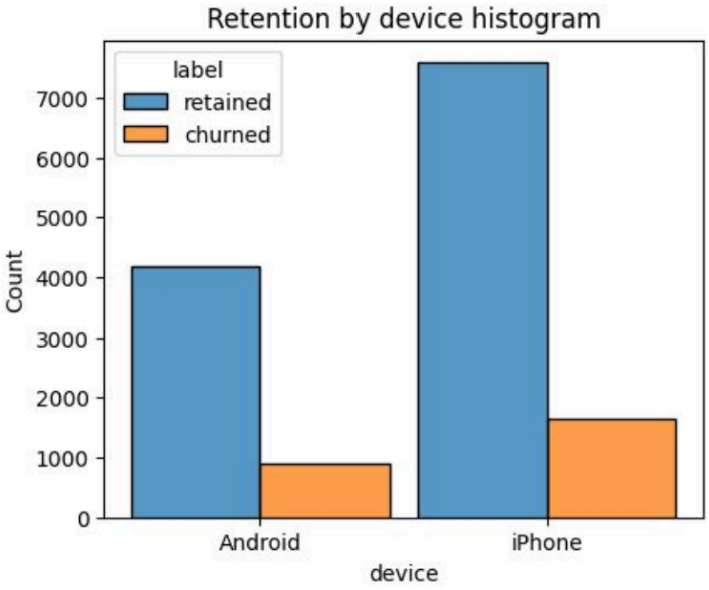
Key Data Insights from EDA (Exploratory Data Analysis)

Exploratory analysis revealed several important patterns in the data:

- **Usage frequency strongly correlates with retention**: 40% of users with zero app usage in the last month churned, while no users who used the app all 30 days churned.

- **Distance per driving day correlates with churn**: Users who drove farther on each driving day were more likely to churn.
- **Tenure was evenly distributed**: Users ranged from brand new to approximately 10 years, with relatively even representation across tenure groups.
- **Data quality issues identified**: Several variables contained highly improbable or impossible outlying values, including `driven_km_drives`, `activity_days`, and `driving_days`.
- **No significant device difference**: A two-sample t-test found no statistically significant difference in the mean number of drives between iPhone and Android users.





Modeling and Evaluation

The project employed a phased modeling approach, progressing from statistical analysis to machine learning:

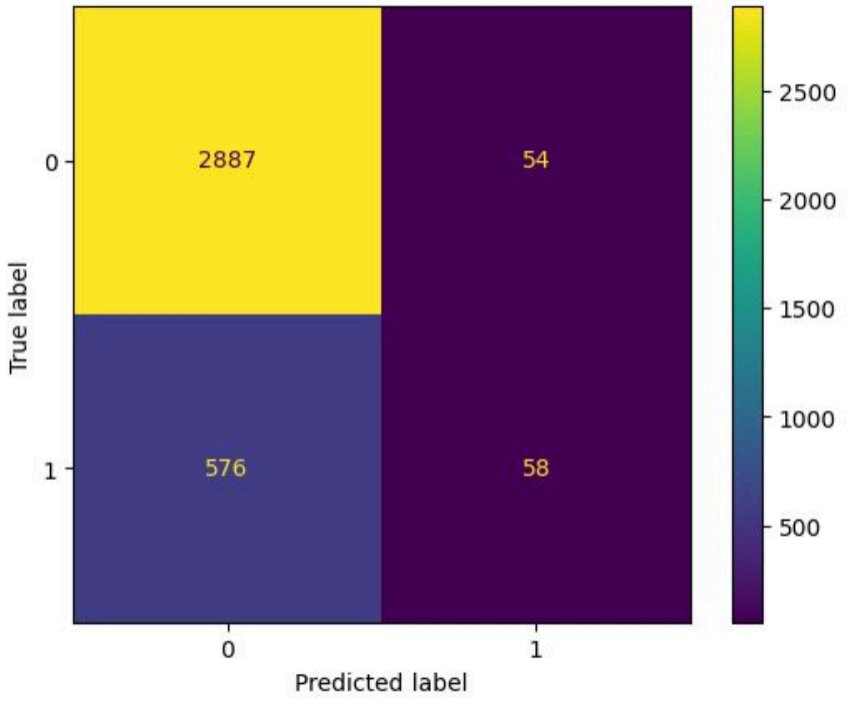
Logistic Regression

A binomial logistic regression model was built as the baseline predictive model. Feature engineering and multicollinearity assessment were performed prior to model training.

Results:

- Precision: 53%
- Recall: 9%
- Most important feature: `activity_days` (negative correlation with churn)

The low recall indicates the model failed to identify most churned users, producing many false negatives.



Note: 1 = churned and 0 = retained

Tree-Based Models

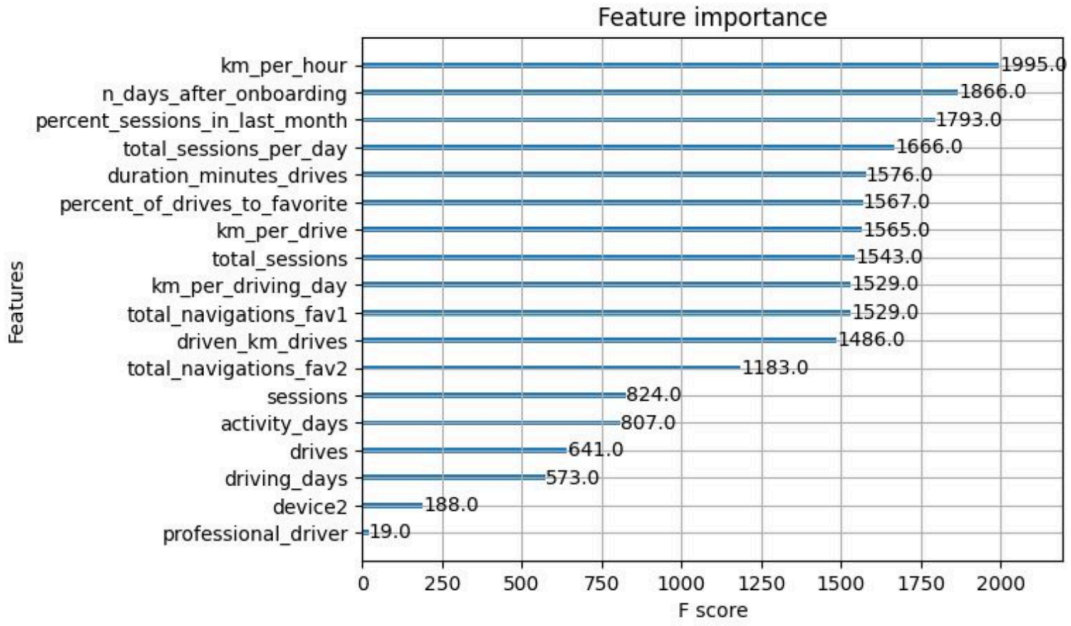
To improve predictive performance, two ensemble models were developed and cross-compared: Random Forest and XGBoost. The data was split into training, validation, and test sets to enable robust model selection and unbiased performance estimation.

Feature Engineering: Six of the top 10 most important features were engineered variables:

- `km_per_hour`
- `percent_sessions_in_last_month`
- `total_sessions_per_day`
- `percent_of_drives_to_favorite`
- `km_per_drive`
- `km_per_driving_day`

Champion Model (XGBoost) Results:

- Recall: 17% (nearly double the logistic regression baseline)
- Maintained comparable accuracy and precision scores
- Outperformed Random Forest across evaluation metrics



The tree-based ensembles offered advantages over logistic regression: higher scores across all evaluation metrics and reduced preprocessing requirements, though with decreased interpretability.

Conclusion

This project demonstrated the challenge of predicting user churn with limited behavioral data. While the XGBoost model improved upon the logistic regression baseline, overall predictive performance remains insufficient for production deployment.

Key Findings:

- User engagement metrics (`activity_days`, `sessions`, `driving_days`) are the strongest churn indicators
- Engineered features significantly improved model performance
- Device type (Android vs. iPhone) has no significant impact on user behavior or churn
- Current data granularity is insufficient for reliable churn prediction

Recommendations:

- Collect drive-level information (drive times, geographic locations, route patterns)
- Capture granular app interaction data (hazard reports, alert confirmations)
- Track monthly counts of unique starting and ending locations
- Pursue a second iteration of the project with enriched data

The insights from this analysis provide a foundation for Waze to refine their data collection strategy and develop more accurate churn prediction capabilities in future iterations.