Waze Project: Monthly User Churn Model

Initial Dataset Exploration and Analysis

Overview

- A dataset consisting of 14,999 observations and 13 variables was provided for initial assessment and exploration.

Objective

- The objective was to determine the suitability of the dataset for further EDA, and whether there were any anomalies or inconsistencies, and whether it can be relied upon to build the predictive model.

Results

- The initial review indicates that there are 700 missing values in the dataset, in the variable “label”, which identifies users as either “churned” or “retained”.
- The review also revealed that churned users seem to drive more times, for longer distances and longer periods of time than retained users.
- When looking at the data per driving day, the churned users also drove an unusual number of times, and completed a high number of km’s, to what would be considered a regular user. The data may represent professional drivers.

Next Steps

- Ascertain the origin of the dataset’s incongruencies, and whether it is suitable for Exploratory Data Analysis and furthermore, the construction of the predictive model.

Waze Project: Monthly User Churn Model

EDA

Overview

- A dataset consisting of 14,999 observations and 13 variables was provided for EDA. The dataset contains 2 string variables, 11 numerical variables, of which 8 are integers and 2 are floats. The dataset also contains 700 missing observations in the label variable.

Objective

- The objective was to work on the initial exploration of the dataset, and further refine through the practice of EDA. This included the checking of proper structure and data formats, as well as obtaining summary statistics, and generating initial visualizations.

Results

- There is a high presence of outliers in many variables and most variables were right skewed.
- The number of drives and sessions have a strong correlation. Retained users have fewer drives than churned users.
- The distance the users drove per day was correlated with user churn. The more they drove, the more they churned. However, the more days they drove, less likely they were to churn.
- There is an inconsistency between the days in activity_days (31) and driving_days (30).
- 18% of users churned, and 82% were retained.

Next Steps

- Ascertain the origin of the dataset's incongruencies, and whether it is suitable for the construction of the predictive model, more specifically the origin and root cause of the missing values needs to be identified, and the discrepancy in days must be resolved.

Waze Project: Monthly User Churn Model

Hypothesis testing

Overview

- The Waze data team is now in the stage of conducting hypothesis testing and attempting to determine if there any key patterns or relationships that can be ascertained as key factors that lead to user churn. This executive summary treats a specific hypothesis raised by the stakeholders.

Objective

- The data team was tasked with conducting a two-sample hypothesis test to determine if there is a statistically significant difference between the mean amount of rides per device type.

Results

- The initial calculations indicate that there is a difference between the mean number of drives per device type. On average: 66 drives for an Android user, and 68 drives for a Apple user.
- However, upon defining the null and alternative hypotheses, choosing a significance level of 5%, and ascertaining the P-value of 14%, by conducting the t-test, we fail to reject the null hypothesis.
- The conclusion reached by the data team is that this difference is most likely due to chance, and does not represent a statistically significant insight.

Next Steps

- As the results of this t-test do not seem to surface any meaningful insight that could help us in our overarching goal of identifying predictors of user churn, we recommend additional testing on other variables.

Waze Project: Monthly User Churn Model

Binomial Logistic Regression Model

Overview

- The Waze data team is now in the construction stage of a Binomial Logistic Regression Model, to see if it is possible to accurately predict user churn while resorting to our known and explore dataset. This executive summary details the construction process and the subsequent results.

Objective

- The data team was tasked with construction a binomial logistic regression model to predict user churn. To that end, feature engineering, model assumption verification, model construction and model evaluation were performed.

Results

- The main insight derived from this process is that the model should not be used to drive business decisions, as it's predicting power is well below a satisfactory level.
- However, it has unearthed the need for additional data, in the sense that more granular features, as to what pertains to how the user utilizes the application, could be of great importance for a more predictive model.
- The conclusion reached by the data team is that the model has mediocre precision (53% of its positive predictions are correct), and very low recall (9% of churned users identified).

Next Steps

- The data team believes that given the model results, there is a need for additional data.
- Furthermore, the results from this model can be used to conduct further exploration and feature engineering.

Waze Project: Monthly User Churn Model

Model Evaluation and Results

Overview

- Waze aims to build a model to predict which users will stop using the app to reduce churn and boost growth. For this purpose, a few potential models were created and evaluated according to key metrics.

Objective

- In order to achieve this objective, the Waze data team developed two different models to compare results and evaluate which model had the highest predictive power: Random Forest and XGBoost.

Results

- Feature engineering demonstrated valuable results, with 6 features being part of the top 10 most predictive features in the model.
- The XGBoost model exhibited stronger scores overall than the Random Forest model. It should be noted that the recall score (17%) is almost double the last logistic regression model built in activity 5.
- It can also be noted that both ensemble models performed better than the singular logistic regression model.

Next Steps

- Given the performance of engineered features. The Waze data team recommends further exploration. It should also be noted that the dataset might not be sufficient for predicting user churn in a reliable manner.