Waze churn prediction
Preliminary data summary

Project Overview

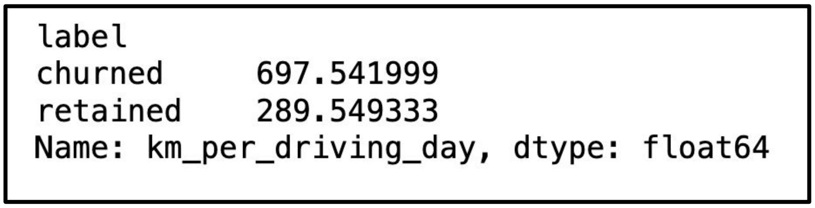
Building a machine learning model to predict user churn on the Waze app. Churn quantifies the number of users who have uninstalled the Waze app or stopped using the app. This project focuses on monthly user churn.

In this part of the project, a preliminary inspection of the data supplied by the Waze was made in order to understand available variables, and ensure the information provided is suitable for generating insights.

Key Insights

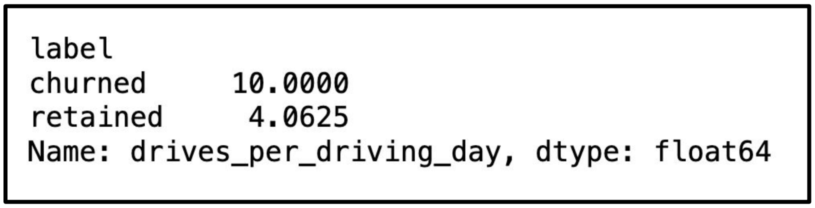
1. This dataset contains 82% retained users and 18% churned users. There are 12 unique variables with types including objects, floats, and integers.
2. Variable 'label' contain 700 missing values (about 4.7% of total entities).
3. 64% of all users were iPhone users, 36% - Android users. The ratio of iPhone users and Android users is consistent between the churned group and the retained group. And those ratios are both consistent with the ratio found in the overall dataset.
4. The median user who churned drove 698 kilometers each day they drove last month, which is almost 240% the per-drive-day distance of retained users. The similar disproportion is for the median number of drives per driving day in the last month.

Details



```
label
churned     697.541999
retained    289.549333
Name: km_per_driving_day, dtype: float64
```

The median kilometers per driving day in the last month



```
label
churned     10.0000
retained     4.0625
Name: drives_per_driving_day, dtype: float64
```

The median number of drives per driving day in the last month

Next Steps

Provide EDA.

Waze churn prediction
Exploratory Data Analysis

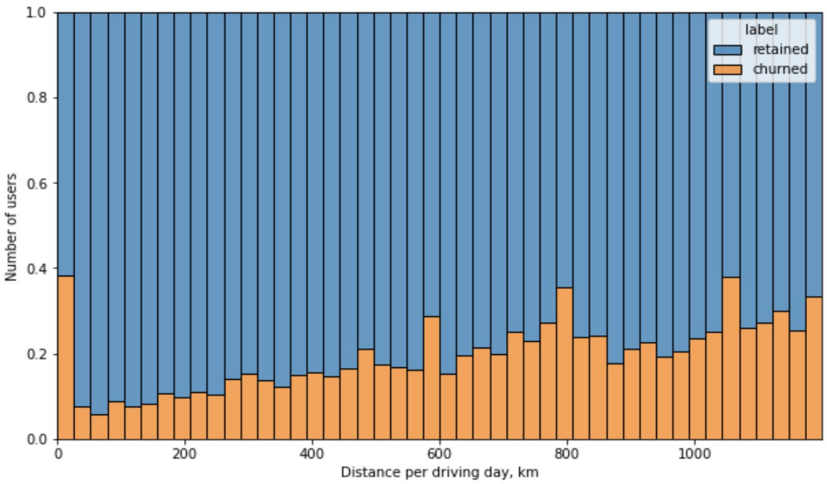
Project Overview

In this part of the project, an Exploratory Data Analysis of the data supplied by the Waze was performed.

Key Insights

1. EDA has revealed that users driving long distances on their driving days are more likely to churn.
2. The churn rate is also higher for people who didn't use Waze much last month. The more times they used the app, the less likely they were to churn.
3. Most of the distributions have a skewed right tail, which indicates outliers in the upper part of the data.
4. About 18 % of users were churned, and 82 % were retained. These numbers are after deleting rows with null values of the 'label' variable.
5. The total user tenure (i.e., the number of days since onboarding) has a uniform distribution, with values ranging from near-zero to ~3,500 (~9.5 years).

Details



The churn rate tends to increase as the mean daily distance driven increases.

Next Steps

- Ask Waze why so many long-time users suddenly used the app so much in the last month.
- Determine the variables that have the largest impact on users churn.

Waze churn prediction
Hypothesis testing

Project Overview

In this part of the project, a two-sample hypothesis test (t-test) was provide in order to analyze the difference in the mean amount of rides between iPhone users and Android users.

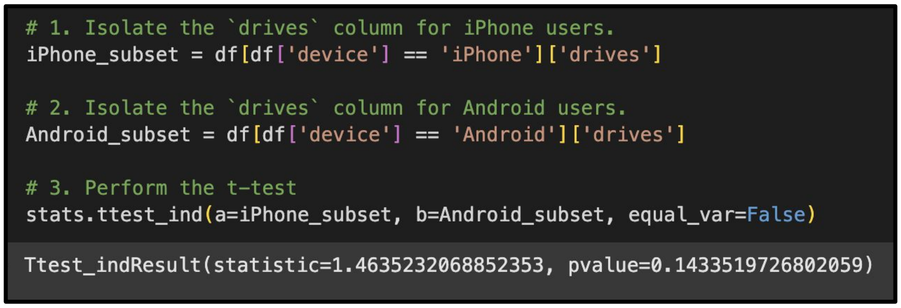
Key Insights

From the result of hypothesis test we can conclude that there is not a statistically significant difference in the average number of drives between iPhone and Android users.

There is no need to improve user experience on a specific device.

Details

H0: There is no difference in the average amount of drives between customers who use iPhones and Android devices.
HA: There is a difference in the average amount of drives between customers who use iPhones and Android devices.



```
# 1. Isolate the `drives` column for iPhone users.
iPhone_subset = df[df['device'] == 'iPhone']['drives']

# 2. Isolate the `drives` column for Android users.
Android_subset = df[df['device'] == 'Android']['drives']

# 3. Perform the t-test
stats.ttest_ind(a=iPhone_subset, b=Android_subset, equal_var=False)
```

```
Ttest_indResult(statistic=1.4635232068852353, pvalue=0.1433519726802059)
```

The P-value is about 14.33%, which is above the significance level of 5%, so we fail to reject the null hypothesis.

Next Steps

Build a regression model to predict user churn.

Waze churn prediction
Regression model

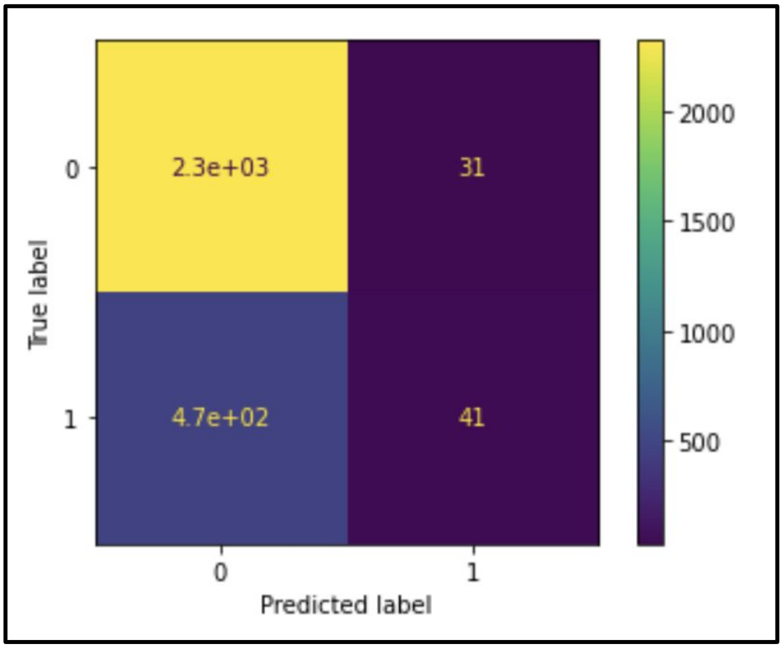
Project Overview

At this stage of the project, a binomial logistic regression model based on a variety of variables was built to predict user churn.

Key Insights

1. The built model uses a range of variables to predict user churn. The variable activity_days most negatively influenced the model's prediction. The more activity days a user has, the less likely he is to be churned. That's not surprising. However, the number of drives unexpectedly influenced the model's prediction: with the rise of this variable, the probability of the certain user being churned rises, too.
2. The model has a decent precision of 56%, meaning 56% predicted as True values (churned users) are True.
3. The model's recall is very low (8%), which means there are many false negative values (it predicts user retention, but this user was actually churned).
4. Most of the variables are not strong predictors of user churn.

Details



Confusion matrix

Next Steps

- The regression model is not recommended for usage to predict user churn.
- Tree-based models should be tried to predict user churn.

Waze churn prediction
Tree-based models

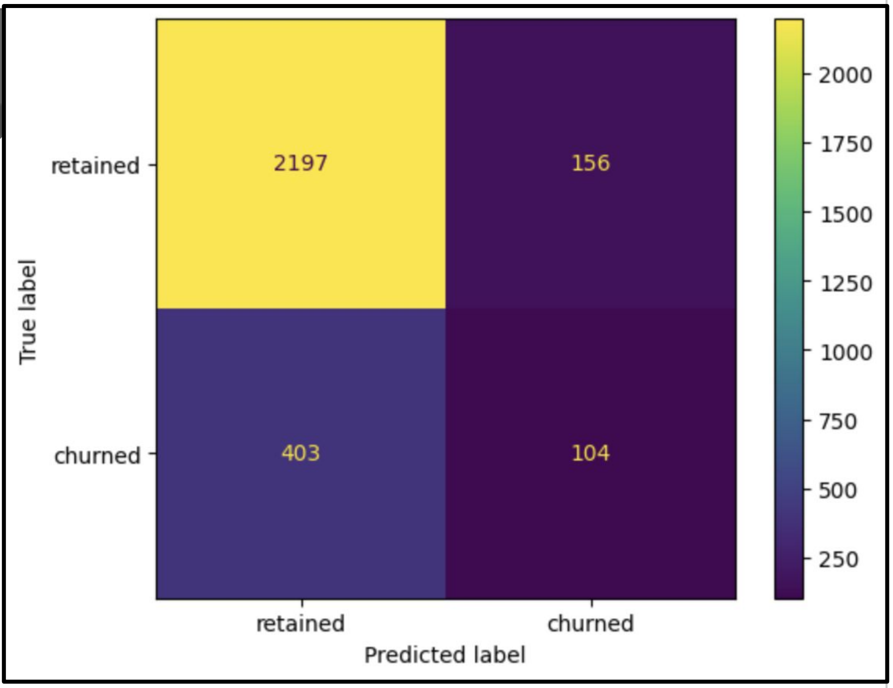
Project Overview

At this stage of the project, two tree-based models were built to predict user churn: random forest and XGBoost.

Key Insights

1. The recall score was used to choose a champion model. Recall explains the percentage of correctly predicted churned users among all actually churned users.
2. On validation data random forest model had a recall score of 12.8%, while XGBoost model - 19.1%. On testing data XGBoost showed even better results - 20,5%.
3. Both models had a decent accuracy (82% and 81% respectively). Since there is class imbalance in data (18% user churned and 82% - retained) this score is unsuitable for model evaluation.
4. The most important features of the XGBoost model were the following features:
   - km_per_hour;
   - number_days_after_onboarding;
   - percent_sessions_in_last_month;
   - duration_minutes_drives;
   - total_sessions_per_day.
5. There are likely to be errors in the initial data. For example, too long driving distance per day, and the number of sessions per month that exceeds the total number of sessions.

Details



Confusion matrix for XGBoost model on testing data

Next Steps

1. Built models should not be used to predict user churn because of the low recall score.
2. Initial data should be checked for errors.
3. Additional data can be gathered to provide new features, such as error report information and users' satisfaction rates.