Waze User Churn Project: Results of binomial logistic regression

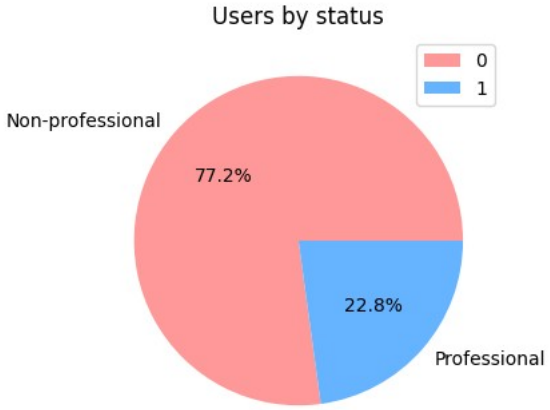
Deliverables

The development of a machine learning model that predicts user churn.
(Since this is an exercise, all models are predetermined. All Python code can be located at: https://github.com/izsolnay/WAZE_Python)

Preliminary conclusions for part IIa

New feature creation

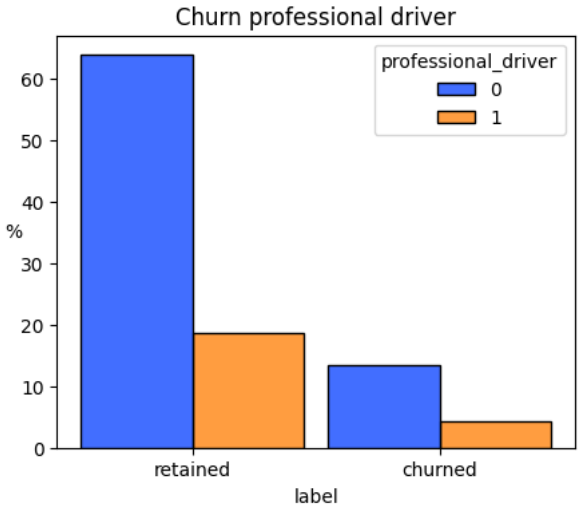
In preparation of the data, a new feature "professional driver" was created as suggested in the previous report. Users who in the last month had either 240 or more drives or who drove over 6000 kms, approximately 3728 miles, were classified as professional drivers. These drivers would have taken at least 8 drives per day or driven roughly 200 kms, 124 miles per day.



Using these metrics, out of the 14,999 WAZE users, roughly 23% qualified as professional drivers. This amounts to 3,265 users of the app.

The median professional driver class doesn't differ much from non-professional class of users in their retention levels, use of fav navigations, activity, and driving days. They did, however, have 24 more sessions per month, took 20 more drives, drove approximately 4,000 more kilometers (~ 2485 miles), and spent 25 hours more on the road.

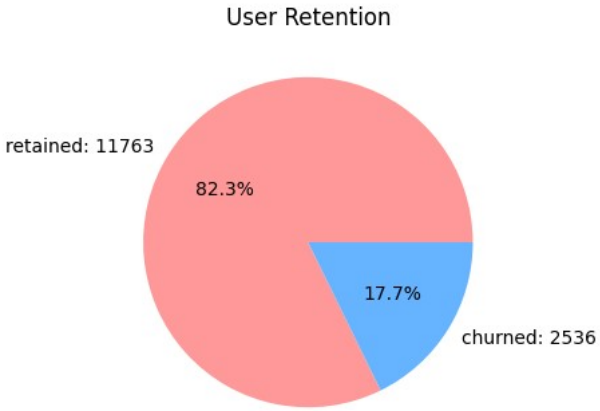
There was also little difference between the percentages of professional drivers who churned and non-professional drivers:



About 17% of non-professional users churned out of all non-professional drivers and 19% of those classified as professional drivers churned out of all users classified as professional.

Results of binomial logistic regression model

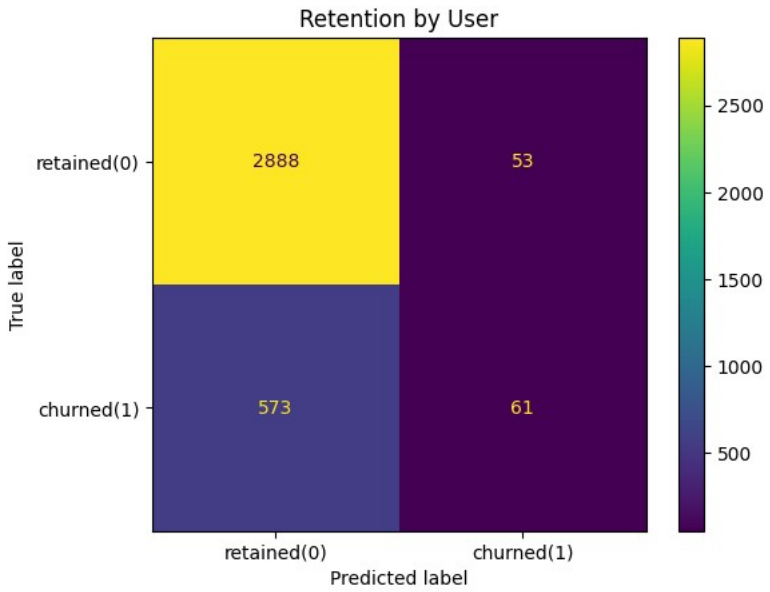
This model used the binary variable 'label', which classified users as either churned or retained, as the dependent variable. Of all users, only about 18% churned (2536), while the remaining 82% (11763) remained loyal to the app:



After checking for correlations between the remaining variables in the data set by generating a Pearson correlation coefficient matrix and checking for linearity of the probabilities with regplots, twelve variables were selected for the regression model. These contained data such as time and distance spent on the road, number of times the app was opened, total sessions, and use of the app's favorite navigations feature.

Test data

Out of the 3575 users in the test data sample 634 churned. Of those, the model captured 61 and missed 573:



Since the goal of this project is to identify whether a user would churn using the available variables, this model's performance is less than stellar. Although it had a near 82% Accuracy rating, its Recall score was 10% with a Precision score of 54%. Because Accuracy is the percentage of correct predictions out of all classifications, included in the score is the model's far better performance at predicting users who would remain with the app.

Accuracy can be misleading if the data set is unbalanced as this one is because it does not account for the distribution of classes (churned/retained). The model could be excellent at predicting one class (e.g., the majority[retained]) and terrible at predicting the other (e.g., the minority[churned]) yet still return a very good Accuracy score. This lack of balance is revealed in the model's f1 score of 16%. The f1 score is the harmonic mean, balancing precision and recall.

Further steps/considerations

- Build, evaluate, and test two decision tree models: random forest and XGBoost
    - Refit models to Recall
    - Consider optimal threshold
- Consider perform a hypothesis test on the new feature 'professional driver' to determine if there is a statistically significant difference in the mean number of rides taken between user classes. Perhaps a campaign targeted toward this class of user would be useful.
- Consider rebuilding logistic regression model with scaled data