Logistic Regression model for Retention Data

Charly Moreno - Data Analyst (Waze's Data Analysis Team)

ISSUE / PROBLEM

It is required to improve the retention strategy of the company to **address the growing churn rate of users from the app**.

To do so, **it was requested to is to build a binomial logistic regression model and evaluate its performance** to answer which could be the main factors impacting churn rate. Our team provides the findings in this report

IMPACT

The model is not ready for use yet. A new version of it, including the 'activity_days' variable, **should be developed.**

The improved version of the model **should aim to improve the recall parameter** since it will be used to detect true positives (true events of churn).

The gathering process of the 'activity_days' variable should be revisited, since it is the most important one (by far), and has an inconsistency with the 'driving_days' variable (when comparing max values). That should be addressed before further modeling.

RESPONSE

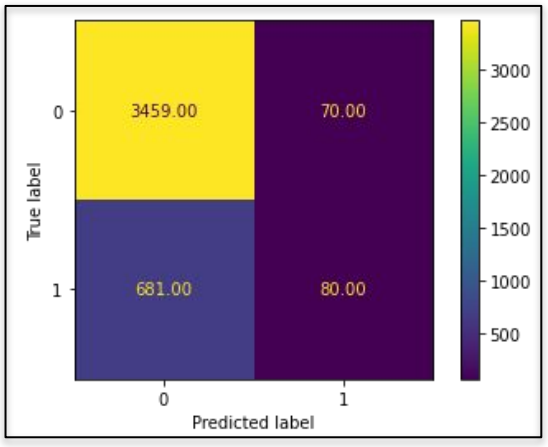
A **binomial logistic regression model was built and evaluated in Python using Jupyter notebook** and relevant libraries (numpy, pandas, sklearn, etc.).

In doing that, **an EDA was also conducted** to make sure the data was correctly gathered and described the business operation that is intended to. Also, to validate if the statistical assumptions to generate the logit model were met.

KEY INSIGHTS

The **precision of the model is not good (0.53,** so 53% of true predictions are correct), and **the recall is very low (0.105,** only 10.5% of churners identified). For that reason, the model should be re-estimated

'activity_days' is the variable that most influenced the model's prediction. In particular, per each new activity day of a user, holding all the other variables constant, its probability of churn decreases by -10%.



1 = Churned user, 0 = Retained user
Activity Days has the most important impact in the prob. of churn

Finally, **another insight is that no other variable had such an impact as 'activity_days'**. The next one is 'drives', with a 0.21% impact of increase in the probability of churn per unit, holding the other variables constant, which also seems inconsistent with the expected sign of the variable.