Relax Inc. Data Science Challenge

Problem:

Defining an *"adopted user"* as a user who *has logged into the product on three separate days in at least one seven-day period,* **identify which factors predict future user adoption.**

Solution:

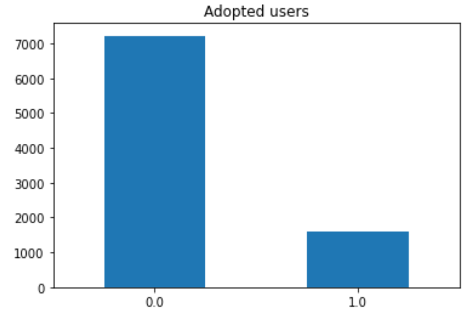
I went thought the following steps in to address the problem:

1. Data Wrangling:

1.1. Data import and inspection: I loaded datasets into Pandas DataFrames and inspected their content.

1.2. Target feature extraction. I extracted an explicit feature indicating if users meet the adopted user requirement:

- I grouped the data by user and day to find which users signed-in at least once daily.
- I unstacked the DataFrames and transposed it to get each user in columns, and days as row indices.
- I applied a rolling window to compute the number of logins in a seven-day window by user and selected the columns with at least three logins. We can see that retained users represent roughly 17% of total users:



Finally, I consolidated both datasets by performing a left join by user id.

1.3. Data cleaning: I removed not valuable information stored in columns and entries (rows)

1.4. Feature engineering: create new features from existing ones that represents more meaningful attributes to predict user's retention. I extracted:

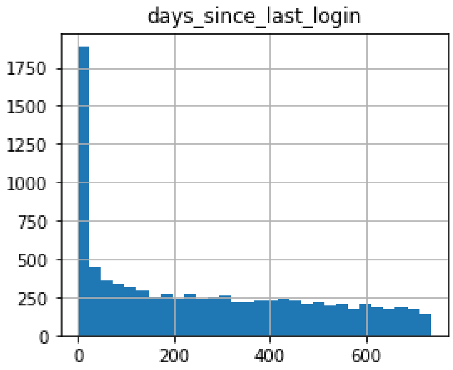
- User domain from email values
- Period of day when the account was created.
- Whether the account creation day was on a weekend or working day
- Age of account
- Days since last login

Later, only the last one (e) proved to be interesting to our problem.

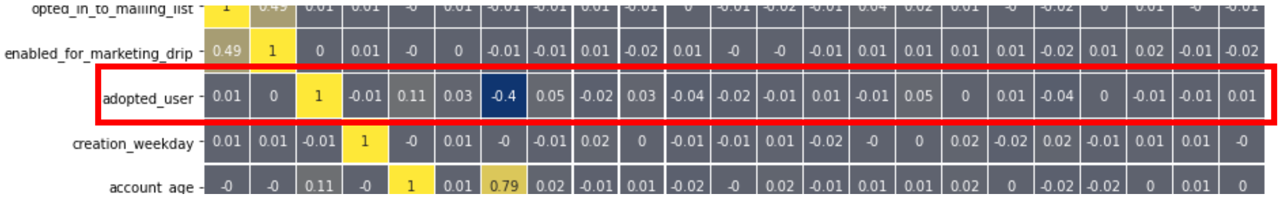
2. Analysis.

2.1. Exploratory data analysis

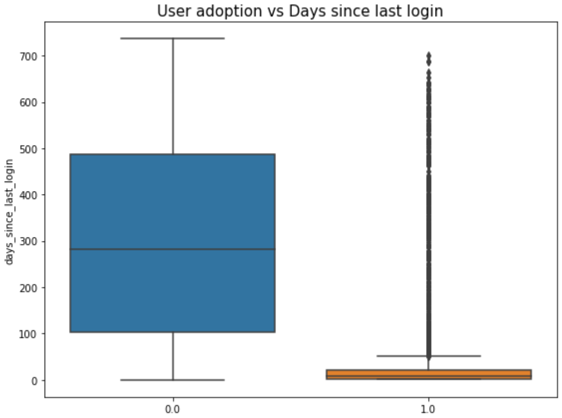
I inspected categorical and numerical features separately to understand their distributions. We can see in the graph below representing days since last login that there is a significant accumulation of users in lower values.



I also explored correlation among features by plotting a heatmap of the correlation matrix. Here is a portion of it:



It can be seen that the there is not any explanatory variable correlated with the response variable, except for days_since_last_login which present a mild inverse association with adopted_user:



We can see from the plot above that adopted users have lower median days since last login than the rest, suggesting that there is a behavioral pattern in adopted users. Also, the distribution of days since last login is more spread out than the rest of users.

3. Findings and conclusions

I have found that user retention is not strongly correlated with any factor contained in the dataset. Though, we could see a moderate association between user's last day of login and our target variable.

Therefore, we can conclude that it is possible to predict user retention by tracking user's recent activity. The more recent user's activity in the app is the higher the likelihood of being a retained user.

I suggest that further studies could focus on finding confounding factors of this association to adjust their impact and increase the correlation between the two variables.

Please, find all code and the rest of figures in this notebook: https://git.io/Jt2cy