Relax Data Science Challenge

Table of contents

[Data](#data)
[Objective](#objective)
[Data Wrangling](#data-wrangling)
[EDA](#eda)
[Modeling](#modeling)
[Conclusion](#conclusion)

Data

For this exercise the following data was provided :

Two CSV files: **takehome_user_engagement**.csv and **takehome_users**.csv
The **takehome_users**.csv has the following columns:

- name: the user's name
- object_id: the user's id
- email: email address
- creation_source: how their account was created
  - PERSONAL_PROJECTS
  - GUEST_INVITE
  - ORG_INVITE
  - SIGNUP
  - SIGNUP_GOOGLE_AUTH
- creation_time: when they created their account
- last_session_creation_time: unix timestamp of last login
- opted_in_to_mailing_list: whether they have opted into receiving marketing emails
- enabled_for_marketing_drip: whether they are on the regular marketing email drip
- org_id: the organization (group of users) they belong to
- invited_by_user_id: which user invited them to join

The **takehome_user_engagement** csv file has the following columns:

- time_stamp
- user_id
- visited

Objective

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. Our target feature is the adopted user.

Data Wrangling

To define the “adopted user” the “timestamp” of the “takehome_user_engagement.csv“ file was resample to a weekly frequency, grouped by user_id and summed the values.
The result was a breakdown of the number of visits per week for each user.
A new data frame was created by grouping the user_id and calculating the maximum number of visits per week and then in a new column “adopted” the a value of 1 was assigned for the users that visited 3 or more times during the week, for the rest was assigned a value of 0.
For the second list the creation time and last session creation time was converted to date time and the difference was calculated to examine the length of time of usage.
The ‘invited by user id’ column showed 45% of missing values so for this reason this column was droped.
The two csv files were merged using the common values of the user_id and object_id.
The categorical data was encoded for further analysis. Two new features were created for better analysis such as “diff” that was the difference of the last Login time minus account creation time.
The emails were separated by companies and enconded “email_com_n”.

EDA (Exploratory Data Analysis)

Several displays were created to analyze the distribution of the different features.

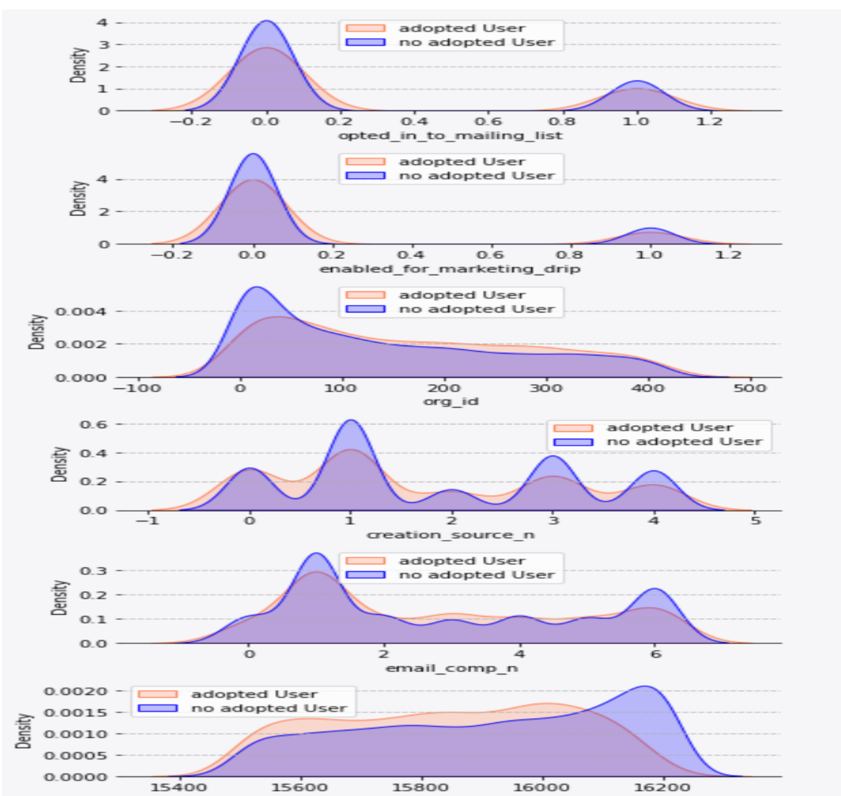

Figure 1: Distribution of the different features.

From the distribution analysis we can see some features that can be useful for prediction such as “time_diff” .
Below is the table of the t-statistics p-values to determine if the features were statistically significant

opted_in_to_mailing_list	pvalue=0.374
enabled_for_marketing_drip	pvalue=0.726
org_id	pvalue=4.455e-11
creation_source	pvalue=0.00019
email_comp	pvalue=0.4233
time_diff	pvalue=9.858e-31

From the table we can see that opted_in_to_mailing_list, enabled_for_marketing_drip and the email_comp aren’t statistically significant so these features were dropped for modeling.

83.6 % were no adopted users and 16.4 % were adopted.
This is an imbalance data set that will require a sample up for the training dataset.

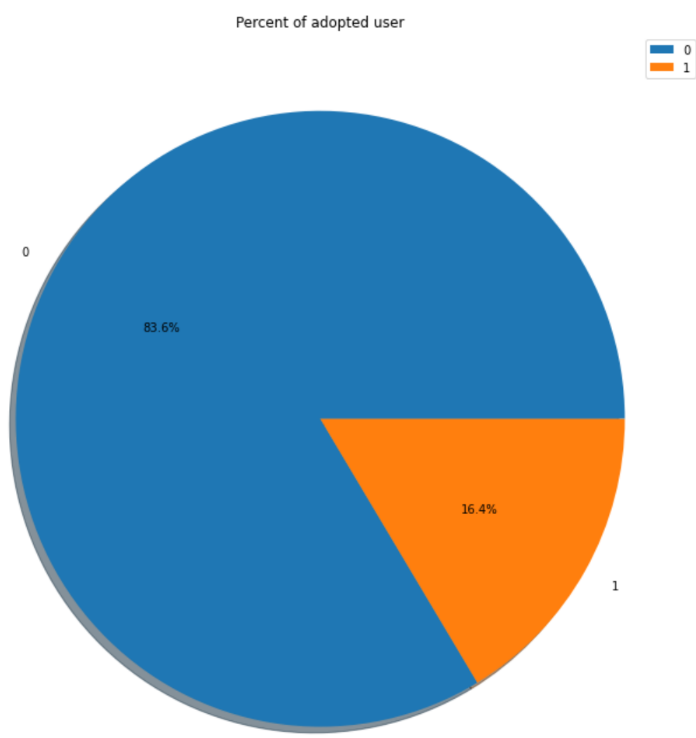

Figure 2: Percentage of adopted users

Modeling

The feature importance showed that the diff of time between the account creation and the last Login ("time_diff") is the main factor for the prediction, followed by organization_id and the creation_source.

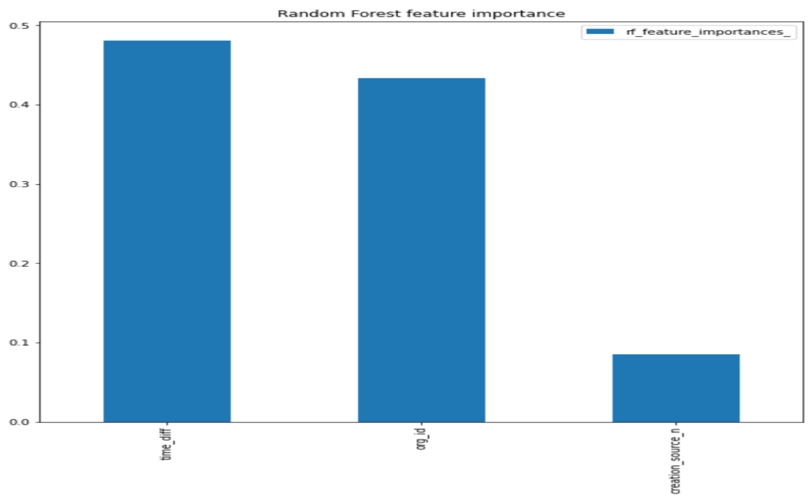

Figure 3: Random Forest feature importance

Two algorithms were tested : Random Forest and logistic Regression. The AUC_score for the two models were close but Logistic Regression was higher as shown below:

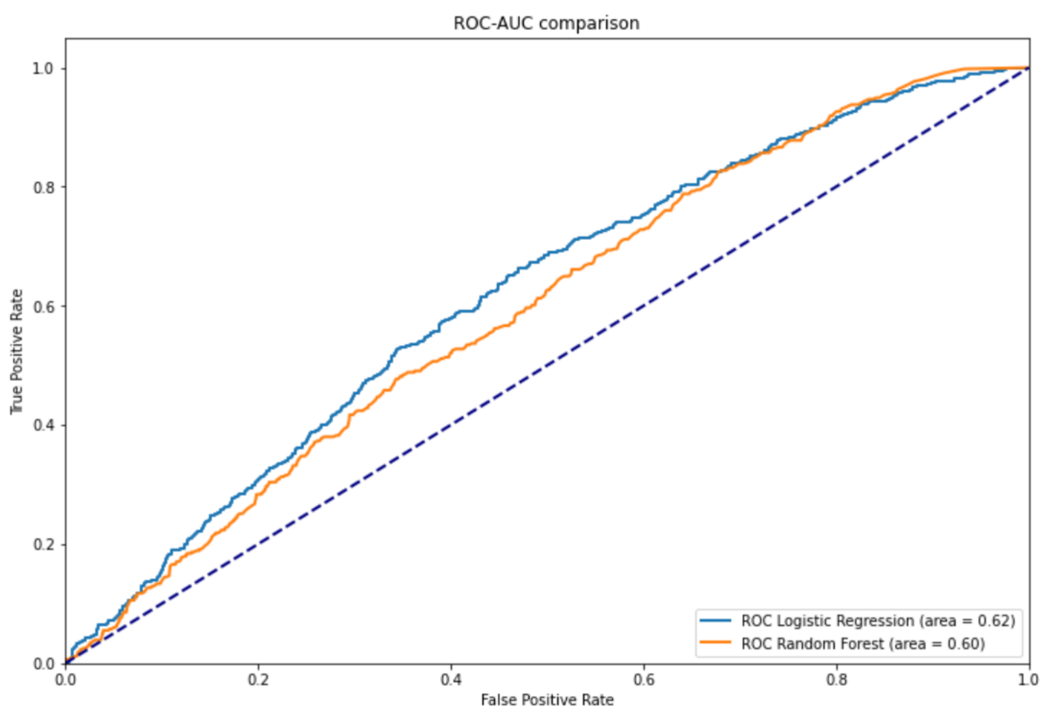

Figure 4: ROC-AUC comparison

Confusion Matrix:

[[1279 934]
[177 257]]

Classification Report

precision recall f1-score support

	precision	recall	f1-score	support
0.0	0.88	0.58	0.70	2213
1.0	0.22	0.59	0.32	434
accuracy			0.58	2647
macro avg	0.55	0.59	0.51	2647
weighted avg	0.77	0.58	0.63	2647

Conclusion

After defining an "adopted user" as a user who has logged into the product on three separate days in at least one seven-day period the data showed to be imbalanced, so it was required to be up-sampled for our training and further modeling. Many of the features were categorical, so for the analysis these features required encoding. Several features were eliminated before modeling because they were statistically insignificant according to the t-test (null Hypothesis). The length of time between the account creation and the last Login showed to be the main factor for the prediction, followed by organization_id and the creation_source. Two algorithms were tested : Random Forest and logistic Regression.

The AUC_score for the two models were close with Logistic Regression having a higher score.
The performance still was not satisfactory (accuracy of 58 percent), consequently I believe this exercise can be tested more for better results, for example analyzing each feature independently.