Marta Taulet
mtaulet@bu.edu

# Feature Importance for adopted users

## 1. Summary

Given a dataset with details about the creation of 12,000 accounts, and a second dataset listing the days in which each user logged in to the product, the goal is to determine the features that make a user more likely to be an "adopted" user in the future. An "adopted" user is defined as a user that logs in to the product at least three times in a seven day period.

## 2. Preprocessing of dataset

In the user description dataset, the "name" and "email" fields were not relevant for the final result as almost all accounts were created by different people with distinct email accounts. Consequently, these fields were dropped. Also, the "last_session_creation_time" did not provide additional information either, as we already had the creation date and time. The same was true for the "visited" field in the user engagement dataset, as it contained a "1" for every entry in the table.

## 3. Labeling

To apply any machine learning method, a labeled column had to be created from the raw data. That is, a tag was needed for each registered account indicating whether they were considered "adopted" users.
Firstly, the accounts with less than three logins were dropped. From the remaining ones, each user was inspected to determine if they had at most seven days between three consecutive logins.The users that passed this test were labeled "1". Otherwise, they were labeled "0".

## 4. Identifying top 5 important features

Two methods were implemented to extract the top features. The first one was to extract scores for feature importances through the use of the SelectKBest scikit-learn python library. It assigned a score to each feature, in this case using the chi-squared statistic, representing how correlated that feature was to the classification label. In other words, it quantified the relationship between each user's characteristic and their adopted/not-adopted class. The second approach was to train a Random Forest classifier (also using the scikit-learn class) and use the method *feature_importances_* to get the desired importance ranking.

## 5. Conclusions

The results from the two methods were similar but not identical [Fig. 0], so the top features for each one were selected and analyzed individually. The most significant fields where: *org_id, creation_day, creation_month, creation_year, invited_by_user_id, opted_in_to_mailing_list, enabled_for_marketing_drip, creation_source, PERSONAL_PROJECTS.*

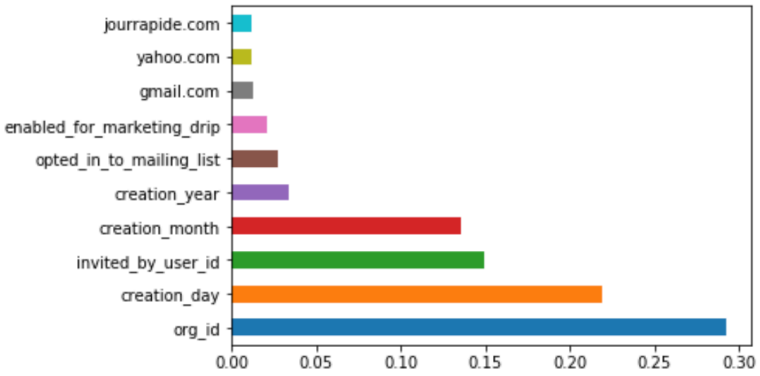
- Even though there were more low *org_ids* subscriptions than high ones, and the distribution had a high variance, in general the probability of accounts with high *org_ids* being "adopted" tended to be higher than low ids [Fig. 1]
- The day of the month had no clear trend but the data suggested a slightly higher probability of "adopted" users the first half of the month [Fig. 2]
- Although May was the month with more accounts created, it was also the month with less "adopted" users, so it is unlikely that May will result in high rate of adoption in the future [Fig. 3]
- The year data was considered not useful because, out of the three years on the record, the beginning of 2012 and the end of 2014 were missing. This led to strongly biased results.
- The feature *invited_by_user_id* seemed to follow a uniform distribution with high variance so the results were not statistically significant [Fig. 4]
- There was a higher chance of "adopted" users when they subscribed to mailing list as well as marketing drip but the difference was more pronounced in the mailing list case [Fig. 5]
- Regarding creation source, guest invites and google signups led to higher adopted users and personal projects had the lowest adoption rates. That would imply that if a future user creates a personal project account the chances of them continuing to use the product for a long time are very low [Fig. 6]

Marta Taulet
mtaulet@bu.edu

**Fig. 0: Top 10 Feature Importance Results**

| | Features | Importance |
|---|---|---|
| 3 | invited_by_user_id | 26272.387638 |
| 2 | org_id | 5418.462606 |
| 6 | PERSONAL_PROJECTS | 56.844560 |
| 17 | creation_month | 36.162874 |
| 18 | creation_day | 20.319532 |
| 4 | GUEST_INVITE | 20.283149 |
| 15 | yahoo.com | 15.237361 |
| 8 | SIGNUP_GOOGLE_AUTH | 13.848040 |
| 12 | hotmail.com | 12.762762 |
| 10 | gmail.com | 11.048830 |

*a. Chi-squared method*




*b. Random Forest method*

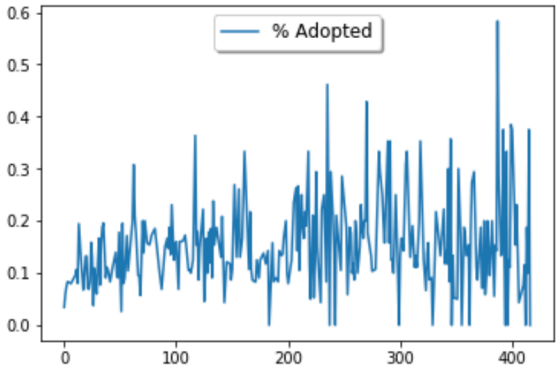



***Fig.1: org_id***

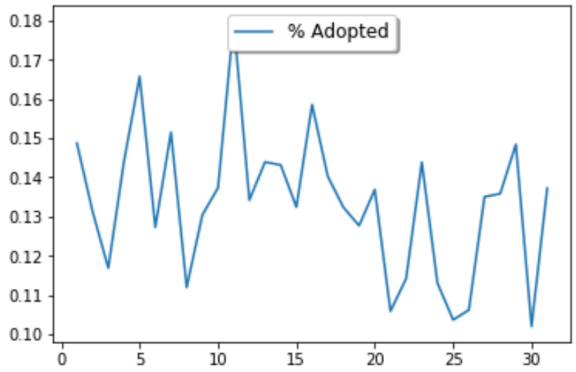



***Fig.2: creation_day***

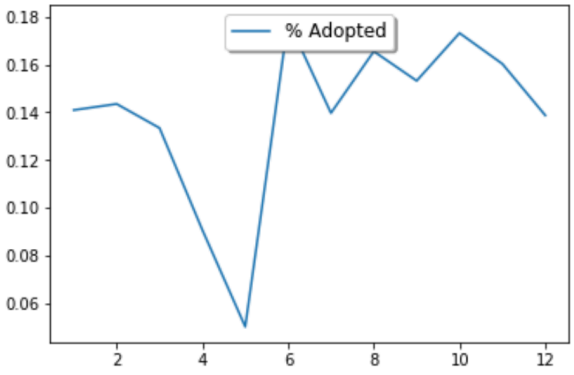



***Fig.3: creation_month***

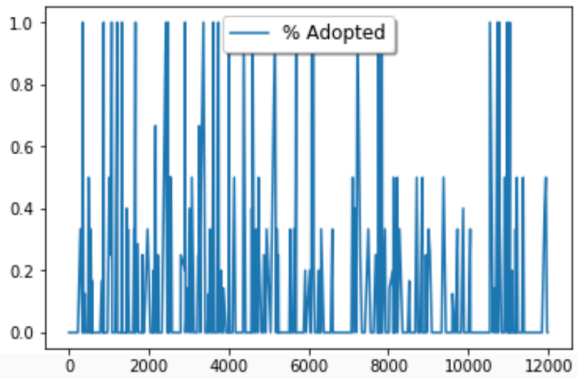



***Fig.4: invited_by_user_id***

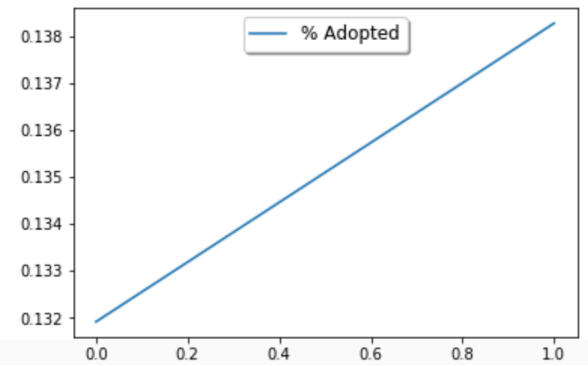



***Fig.5: opted_in_to_mailing_list***

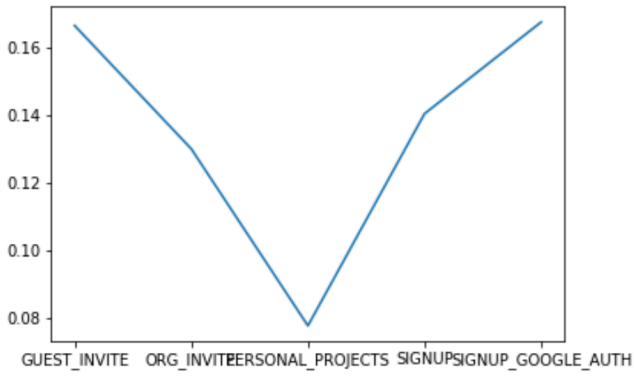



***Fig.6: creation_source***