Identifying Adopted Users Factors

To identify correlated factors of future user adoption, I built two predictive models: Extreme Gradient Boosting (XGBoost) and Logistic Regression. XGBoost can learn non-linear relationship between factors and target, and it allows us to retrieve information of how import each factor is to predict the target. Logistic Regression picks up linear correlation between factors and target, and we can use its coefficients to assess factor importance. Using multiple models can help better verify the effects of our factors on the target. By combining factor importance from diverse models, we can have a better understanding on how these factors affect future adoption. Beyond selecting proper models, feature engineering always plays an essential role in the performance of our models and how the questions are answered. In this project, I created many new factors and four of them have significant effects on predicting future adoption: login frequency, duration, organization size and email domain. The details of these factors are explained in the following paragraphs.

Login frequency is the dominate factor of predicting future user adoption in both models. It is extracted from the usage summary table by calculating how many times users had logged into the product until the data were collected. As shown by the heatmap in the following page, login frequency is highly correlated with user adoption. It makes sense because as the frequency of users logged into product increases, they are more likely to become adopted. This factor provides us important information that if we are able to attract users to the product, we are likely to increase the number of adopted users. For instance, we can provide users friendly tutorials of the product or make the product easier to use.

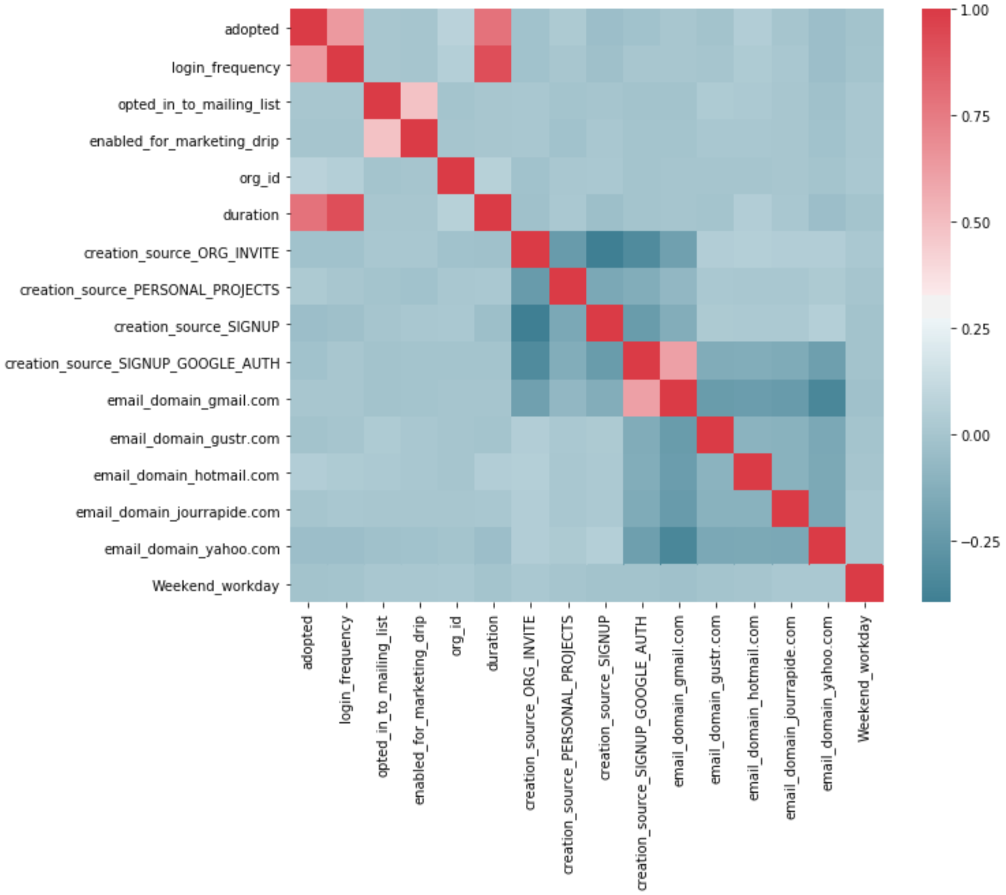
The second important factor for predicting future adoption is the duration. It is calculated by subtracting the time that users created their account from the last time they logged into the product. A non-adopted user with large duration indicates that the user is a potential adopted user in the future. Getting user into the mailing list might helpful for increasing the duration because it keeps reminding users of Asana’s product and attracts them to check the new version product.

The third factor that is essential to future adoption is the current number of users of a specific organization that is using Asana’s product. It is calculated by counting the times of a specific organization ID that appears in the dataset. This makes sense because more people in an organization using the same product, they are more likely to keep using it.

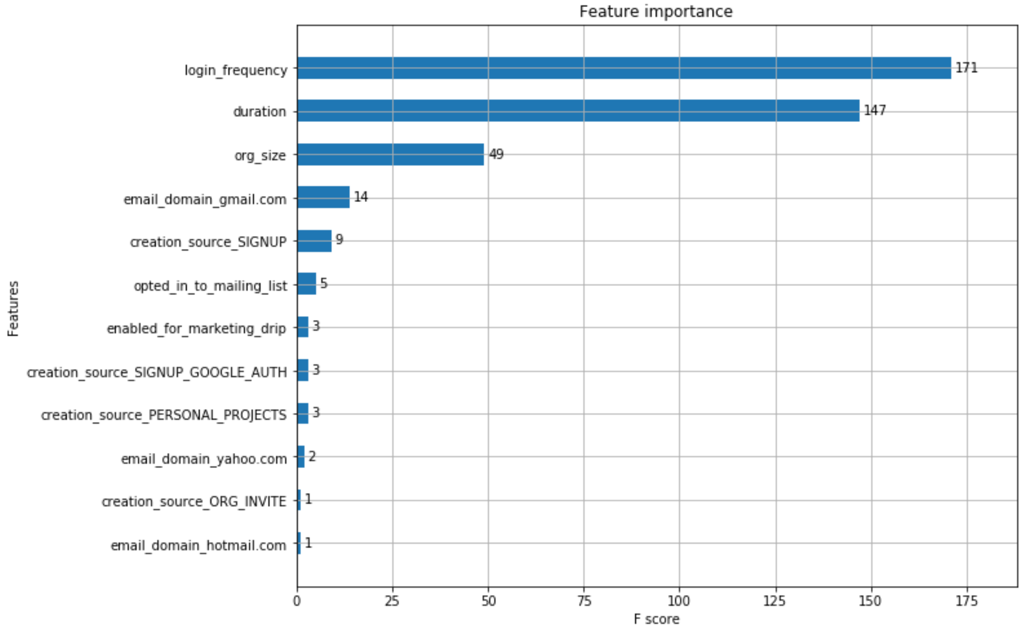
The models also show that users with gmail are more likely to become adopted users. By data exploration, I found out that the reason why gamil domain is important is that it has more users. Hotmail.com and yahoo.com are also important indicators. As for email domains, most of them are disposable emails. Disposable emails are only allowed to receive emails, which means they are not formal emails. Therefore, they are either not important factors or play a negative role in future adoptions.

In conclusion, both models have nearly 100 percent accuracy of predicting adoption. Therefore, the factor importance we retrieved from the models are pretty convincing. One drawback of the modeling is that it tends to predict new users as non-adopted users, because the duration and login frequency is low for new users.

Correlation of each factors with adopted users



Ranking of factor importance


Feature importance
Features
login_frequency
duration
org_size
email_domain_gmail.com
creation_source_SIGNUP
opted_in_to_mailing_list
enabled_for_marketing_drip
creation_source_SIGNUP_GOOGLE_AUTH
creation_source_PERSONAL_PROJECTS
email_domain_yahoo.com
creation_source_ORG_INVITE
email_domain_hotmail.com
F score