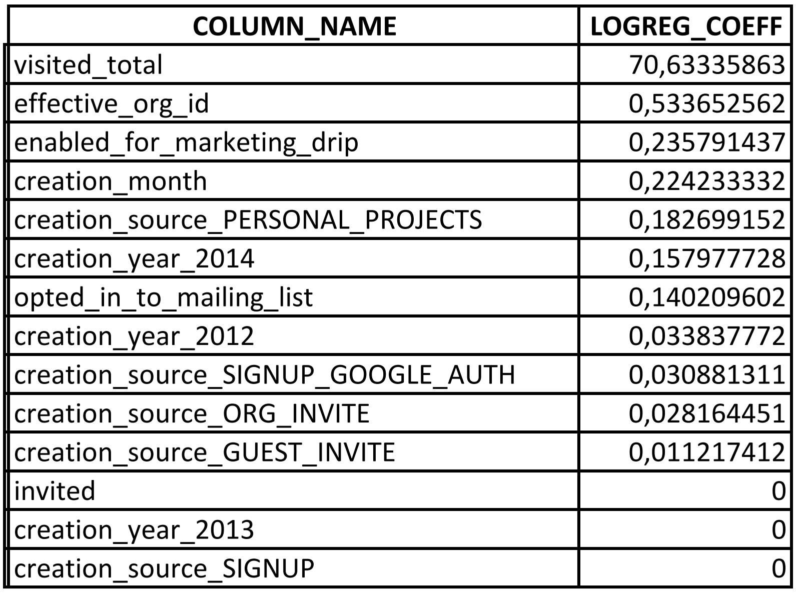
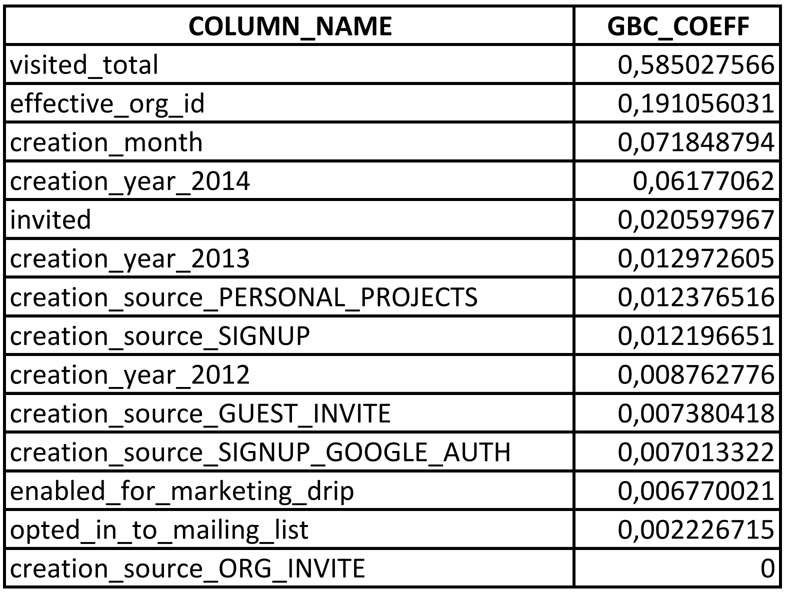
Relax Data Science Challenge: Results

To determine which factors predict future user adoption we did the predictive modelling using the Logistic Regression and Gradient Boosting Classifier methods. The predictor variables (13 in total), generated based on the columns in two provided datasets, are: (i) ‘visited_total’ describing the total number of logins for each user, (ii) ‘effective_org_id’ describing the percentage of users who became adopted for a given organization, (iii) ‘invited’ – the binary feature telling us whether the user was invited, (iv) ‘opted_in_to_mailing_list’ and ‘enabled_for_marketing_drip’ that have the same meaning that is given in the instructions, (v) ‘creation_month’ giving the month in which the account was created, (vi) three binary features ‘creation_year_2012’, ‘creation_year_2013’ and ‘creation_year_2015’ telling us the year in which the account was created, (vii) five binary features ‘creation_source_PERSONAL_PROJECTS’, ‘creation_source_GUEST_INVITE’, ‘creation_source_ORG_INVITE’, ‘creation_source_SIGNUP_GOOGLE_AUTH’ and ‘creation_source_SIGNUP’ that describe the source of account creation as described in instructions.

Two prediction methods that are of different nature give accuracies that do not differ much, and both accuracies are high (around 99 percent). Two methods lead to different relations between precision and recall. The first table below provides the relative importance of features obtained from Logistic Regression, while the second one summarizes the results from the Gradient Boosting Classifier.



COLUMN_NAME	LOGREG_COEFF
visited_total	70,63335863
effective_org_id	0,533652562
enabled_for_marketing_drip	0,235791437
creation_month	0,224233332
creation_source_PERSONAL_PROJECTS	0,182699152
creation_year_2014	0,157977728
opted_in_to_mailing_list	0,140209602
creation_year_2012	0,033837772
creation_source_SIGNUP_GOOGLE_AUTH	0,030881311
creation_source_ORG_INVITE	0,028164451
creation_source_GUEST_INVITE	0,011217412
invited	0
creation_year_2013	0
creation_source_SIGNUP	0



COLUMN_NAME	GBC_COEFF
visited_total	0,585027566
effective_org_id	0,191056031
creation_month	0,071848794
creation_year_2014	0,06177062
invited	0,020597967
creation_year_2013	0,012972605
creation_source_PERSONAL_PROJECTS	0,012376516
creation_source_SIGNUP	0,012196651
creation_year_2012	0,008762776
creation_source_GUEST_INVITE	0,007380418
creation_source_SIGNUP_GOOGLE_AUTH	0,007013322
enabled_for_marketing_drip	0,006770021
opted_in_to_mailing_list	0,002226715
creation_source_ORG_INVITE	0

Looking at the feature importance, we can make the following conclusions.

1) In all methods the most important feature is the total number of logins. The more visits the user made, the higher probability that the user became an adopted one. This result is quite natural but does not answer the question which other factors stimulated logins.

2) The second most important feature is 'effective_org_id' that gives the percentage of adopted users affiliated with a given organization. There are organizations with fraction of adopted users as high as 0.3-0.5; however, there are organizations with extremely small fraction of adopted users. In order to stimulate logins, and hence increase the number of adopted users, it is useful to understand what the interests of organizations with high and low percentages of adopted users are, and what the affiliated users are looking for. One needs to analyze other datasets with the information about the specific purposes of logins.

3) It is interesting that the month when the account was created plays a rather important role in both methods. For some reasons, users who created their accounts in April and May have the lowest fractions of adopted users among them. At the same time, the number of accounts appeared in May is the largest.

4) The binary feature that tells us whether a given user was invited to join another user's personal workspace or not, plays the highest role among the features related to the source of account creation. The users who were invited to join another user are least likely to become adopted users.

5) Finally, one more important feature in both methods is whether the account was created in 2014 or not. 2014 is the latest year, and it is important to pay closer attention to other possibly relevant datasets related to this year.