Relax Data Science Take Home Challenge

Summary, Results and Recommendations by Garrick Chu (April 2018)

Given the data provided, most of the effort was focused on cleaning and wrangling the data. This was required to correctly identify "adopted users" and merge with user data to then be assessed as predictive factors.

From the users table, I excluded data series which I believe to have no bearing as a predictive factor including name, email, and creation_time. Furthermore, I used one-hot encoding of the 'creation_source' field to facilitate modeling.

With the data processed, I utilized a function from SciKit-Learn (a popular Machine Learning library for Python) called Recursive Feature Elimination (RFE). The basis of this function is that it continually tests different variations of a model using various feature weightings and determines the best. Here, I used RFE on a Logistic Regression model (which outputs prediction of either a 1 or 0, "adopted user" or "not adopted"). Using RFE, we arrive at the below results of possible predictive factors:

	Feature_Rank	Feature_Selected
max_weekly_visits	1	True
creation_source_GUEST_INVITE	1	True
creation_source_ORG_INVITE	1	True
creation_source_PERSONAL_PROJECTS	1	True
creation_source_SIGNUP	1	True
creation_source_SIGNUP_GOOGLE_AUTH	2	False
opted_in_to_mailing_list	3	False
enabled_for_marketing_drip	4	False
org_id	5	False

The above table suggests that a user's sign-up platform and permission-type are predictive factors in whether a user will continue to be engaged. The implication is that adoption rates can be influenced and improved by improving the product wherever we determine the 'creation_source' to be least accretive to adoption rates. In other words, if Limited-permission users are not as engaged, then we can implement strategies to improve the UX for these users to drive engagement higher.

Code can be found on Github: https://github.com/thegarrickchu/Springboard-TakeHomeChallenges/blob/master/Relax_Inc/Relax_Data_Science_TakeHomeChallenge_GC.ipynb