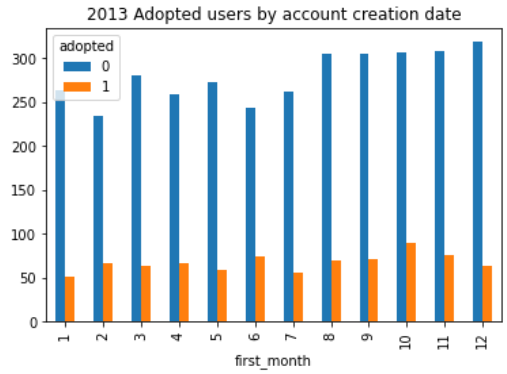
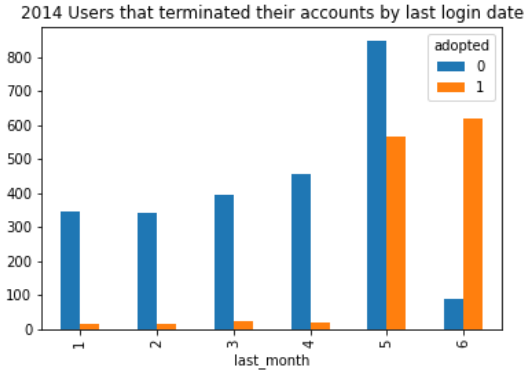
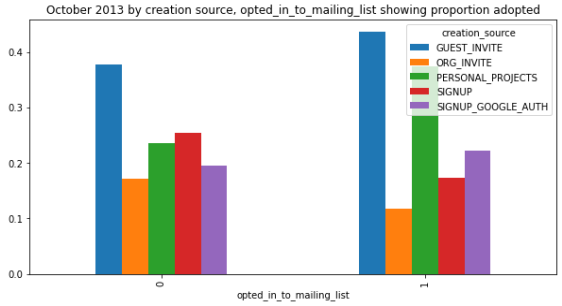
Relax Inc Take Home Challenge - Factors for Predicting Future User Adoption

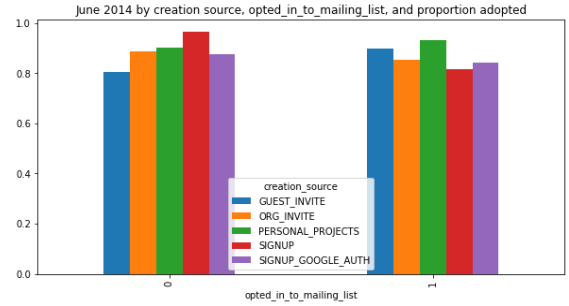
By using the client's definition of adopted users (logged in at least 3 times in a 7 day period), it was found that the highest adoption period for new accounts was in Oct 2013, also, the highest adopted user account termination period was June 2014. On the graphs below the month is depicted by its number, note that because of the difference in distributions (note differences between adopted 1 and not adopted 0 users), these periods were considered highly important for finding useful features for prediction.



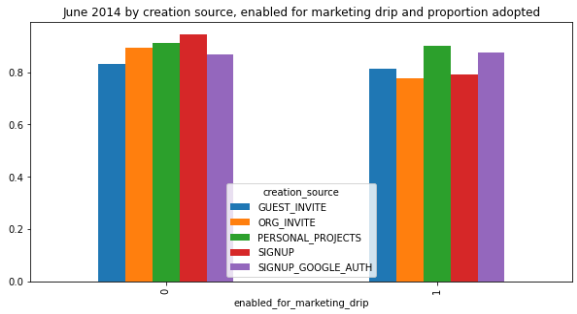


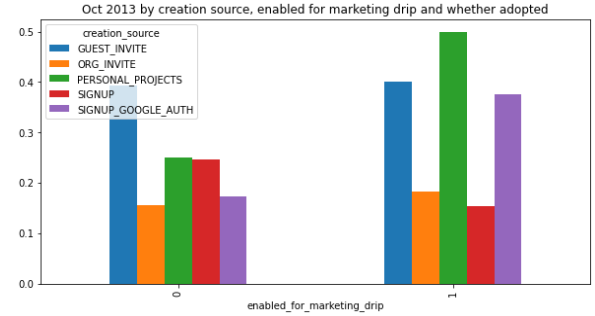
Zooming into October 2013 shows the highest proportion of adopted users invited through guest invites and opted into marketing emails (mailing list) vs adopted users who terminated their accounts in June 2014 -- most of these users in contrast signed up via the website (signup) and did not opt into marketing emails as seen on the graphs below.





Those adopted users who terminated their accounts (June 2014) had the highest proportion not enabled for marketing drip (on regular marketing emails). In contrast, the highest proportion of new, adopted users (Oct 2013) had their accounts enabled for marketing drip and signed up through personal projects (they were invited to join another user's personal workspace) as seen here.





Adopted users were associated with longer active time periods, those who terminated their accounts even more so, this is best seen through the active_days feature, the plot showing active days vs adopted and not adopted is very large and more easily seen in the Notebook. Finally org_id (user's group) was very good at separating adopted vs not adopted users. Based on the results of EDA, expected feature importances would be: active_days (derived from time features), time features, org_id, provider (derived from email), GUEST_INVITE, opted_in_to_mailing_list, enabled_for_marketing_drip, SIGN_UP, PERSONAL_PROJECTS. A Random Forest Classifier confirmed this intuition, although, with some minor changes. Metrics obtained from testing on the adopted class: precision: 0.84, recall: 0.93, f1-score: 0.88. Top ten features in order of importance as determined by the classifier: active days, time features, org_id, invited_by_user_id, ORG_INVITE, provider, opted_in_to_mailing_list, SIGN_UP, enabled_for_marketing_drip, SIGNUP_GOOGLE_AUTH. It can be seen that ORG_INVITE would help explain variance very well, just visually as seen above, it separates the dependent variable the most between the two times, so that is not a surprise, neither is SIGN_GOOGLE_AUTH. The surprise here is invited_by_user_id (Notebook), because it just didn't look like it was grouping the dependent variable well enough, to be frank, it was only tested as experimentation rather than being chosen from EDA, but it came out on the top 10 features.