Dao Vang
dao-v.github.io
Repository: Predicting_High_Usage_Customers

Challenge Summary: Relax Inc.

The purpose of this challenge was to examine and use the two datasets provided to predict if a customer would be labeled as an 'adopted user', 'a user who has logged into the product on three separate days in at least one seven day period', with the focus on finding what factors/variables about the user could impact whether or not they would be considered an adopted user (user adoption). In the script used to perform the computation, I interpreted the definition of an adopted user to be a user who has logged in three separate days within a week. Two random forest models generated produced a bi-categorical prediction accuracy of ~98% and ~86%, depending on if the total number of times a user logged in was included.

When examining the feature importance generated from random forest models, the most important features were the total number of logins for a user (if provided) and/or the method the account was created ("creation_source") (Table 1). Of the methods used to create accounts who were labeled as adopted users, ~34.7% were adopted users who were sent invitations by an organization (Table 2). However, these ~34.7% were invited as users with "full members". If the other users who were sent invitations by organizations who would only have limited permissions (guests), the total percentage of users associated with organizations would be ~56.9%. In contrast, personal users and others associated with personal users resulted in a total of ~43.1% of users. (To see what other types of methods were available, the definitions of these methods, and what data each dataset contained, see the PDF in the repository named "relax_data_science_challenge.pdf".)

An odd correlation found was that none of the personal users were invited by another user, even though the definition of one of the methods for account creation was defined to require an invitation to join another personal user's workspace. This correlates with the total percentages in the previous paragraph; users who were not invited by another to be ~43.1% and invited users to be ~56.9%.

If recommendations were wanted from Relax Inc. on what factors they could focus on when wanting to increase usage or attract specific customers, I would suggest that they focus on business-to-business marketing and on individual personal users. Because the users are almost split between organization users and personal users, it is better to target the source of these users. For example, it is better to target management-level employees and ignore employees who cannot make decisions on user adoption in the organization. Another example could be to market towards individual users at conventions because individual users will indirectly bring other users into their projects.

If Relax Inc. wanted to use either of the two random forest models generated in the script, they should only use one model depending on the question to be answered. If the question is if a new user will become an adopted user in the future, using the model trained without the total number of logins for that user should be used to help with predicting. If the question is if a current user becomes an adopted user in the future, then the model trained with the total logins for that user should be used.

In conclusion, either model produced was good for predicting consumer usage. However, to better business recommendations, more information on how users use the program and who they work with (if any) would better improve understanding the relationship dynamics between connected users.

Table 1: Feature Importance (in decimal format for percentage)

	creation_source	opted_in_to_mailing_list	enabled_for_marketing_drip	invited_by_user_id	TOTAL_LOGINS	ACCURACY
WITH TOTAL	0.007957	0.002636	0.002939	0.001335	0.985133	0.983000
WITHOUT TOTAL	0.747783	0.103724	0.091756	0.056737	NaN	0.863667

Table 2: Method of Account Creation (percentage of adopted users)

PERSONAL_PROJECTS	GUEST_INVITE	ORG_INVITE	SIGNUP	SIGNUP_GOOGLE_AUTH
0.103865	0.222826	0.346618	0.182367	0.144324