Relax, Inc. Takehome Challenge

Cleaning

The dataset included many missing values for last sign-in timestamp and invited_by_user_id. Last sign-in timestamp was imputed as the account creation timestamp, after confirming that those users didn't appear in the login dataset; these users were then set aside as inactive during analysis. Invited_by_user_id was imputed as a new user_id value, 0, to indicate that no user invited this user to join. Object features became categorical features.

Wrangling

To identify adopted users, the login time series was grouped by username and resampled at a seven-day interval. A count of unique timestamps within each seven-day interval yielded a list of adopted users (users who had logged in at least three times in at least one seven-day period).

Feature Architecture

I derived the account creation date's day of the week, and I created a new feature for whether or not the user invited themselves to join.

Exploration

A count plot of account creation source reveals that organizational and guest invites yield adopted usership, and a breakdown of the number of adopted users per organization revealed many orgs with low numbers and a small number of orgs with high numbers. (See this repo's plots directory for these plots.) Personal projects users were a bit more likely to open an account on a Friday.

Modeling

Class imbalance led to a random forest classification model with high accuracy but low precision and recall scores, due to the relatively small number of positive labels in the dataset; class balancing improved performance slightly. The model's feature importance emphasized marketing (mailing list and marketing drip) foremost, followed by creation source. Likewise, K-Means clustering yielded models that segregated primarily according to account creation type, emphasizing the importance of this feature in considering customer segmentation.

Further Research

It might be worth looking into specific practices at the small number of organizations with 4-8 adopted users to see if those might be shared more widely toward increased conversion.