Relax Inc. Take-Home Challenge

Rachel Chiang, 2023-08-02

Guidelines: Define an “adopted user” as a user who has logged into the product on three separate days in at least one seven-day period, and identify which factors predict future user adoption.

Out of the 12000 users, only 1656 of them were found to be `adopted`, or about 13.8%, so our dataset was moderately imbalanced. Appendix, Figure 1 depicts a correlation heatmap, and the response variable, `adopted`, has a very slight positive correlation with `org_id` and `creation_source`.

To begin with, I quickly tried to predict `adopted` using logistic regression, dropping the dates and unique features (such as email). Unsurprisingly, it failed to predict any positive labels, so I resampled the dataset by a factor of 6, thus ending up with 1656 positive and 1724 negative labels. At the cost of accuracy, the model now made some positive predictions (Appendix, Figure 2).

Using GridSearchCV, I tried some other hyperparameters for Logistic Regression and for Random Forest Classifier models. The Random Forest found comparable results, but the first Logistic Regressor on the downsampled dataset had the best predictions.

In agreement to the correlation heatmap, the most important feature according to the model was the `org_id`, and then it found SIGNUP_GOOGLE_AUTH (`creation_source`) to be the next most significant. From my understanding of the `org_id`, it could make sense because it means that there are probably work-related, social, or educational incentives that cause users to log in often. As for the `creation_source`, I do not completely understand what it means, but I speculate that having a convenient way to sign up and sign in is important for the user to sign up and continue signing in in the first place.

The model was not very accurate, as can be seen in the confusion matrix attached below. Perhaps a future work could be to accommodate the class imbalance more. As examples, the class weights or the threshold for the probability to be considered a positive or negative label could be tuned. Completely different models could also be tested because I only tried two relatively basic ones. Otherwise, more features could be engineered, since only the present data was used. A few examples include generalizing the `org_id`, `invited_by_user_id`, and the `creation_source`, since the distributions are a bit skewed and maybe we don’t care about who specifically invited whom (or at least we don’t have much information on the characteristics of the users themselves, though understanding our users might be helpful too). Perhaps our dates could be transformed into behavioral information, such as login frequencies or raw total logins or times of logins. What has been done in the project is only a start.

Github link: https://github.com/raechiang/Springboard/tree/main/Unit_27-2-3_Relax_Take-Home_Challenge

APPENDIX

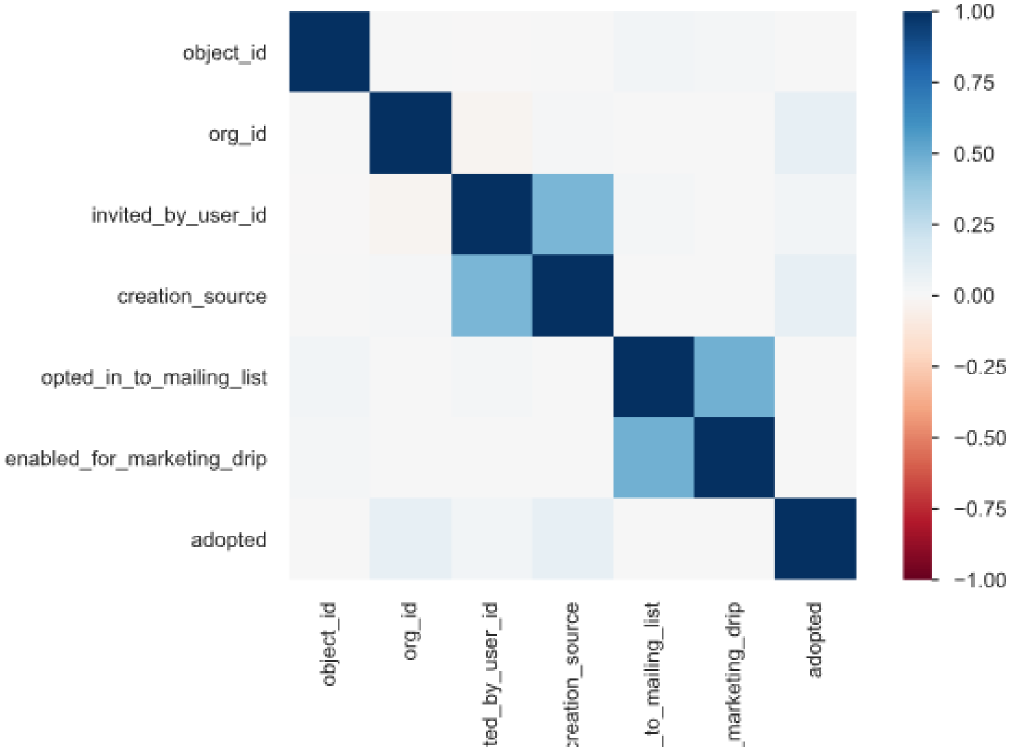

Figure 1. Correlation heatmap for user data.

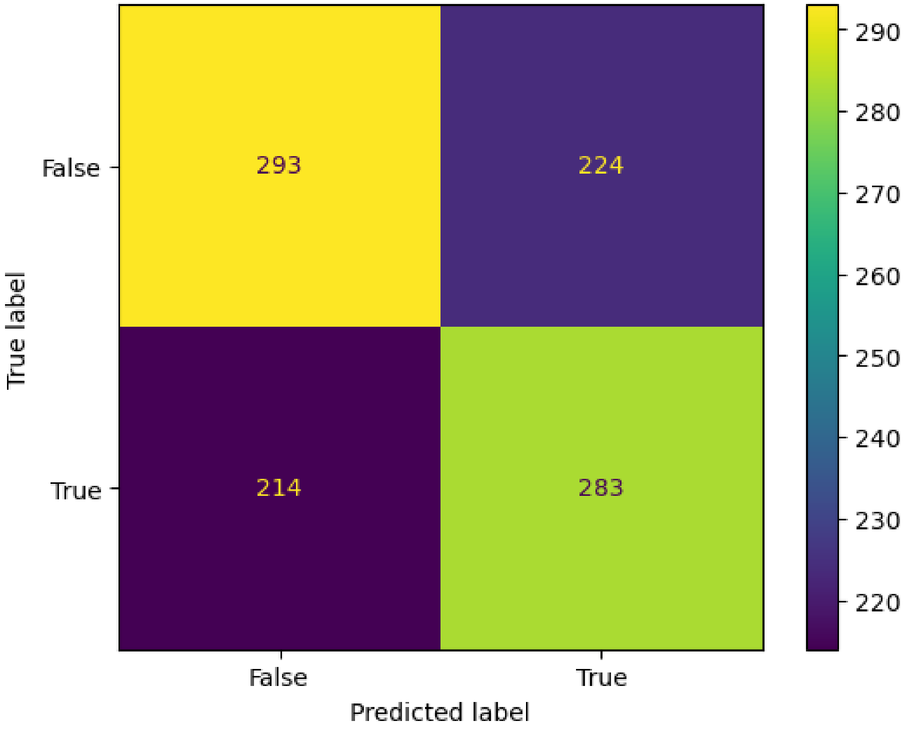

Figure 2. Confusion matrix for Logistic Regression with a downsampled dataset (51/49%).