Relax Challenge Write Up

Referenced code can be found here:
https://github.com/kjd999/Springboard-files/blob/master/Relax%20Take%20Home%20Challenge/Relax%20challenge.ipynb

The purpose of this project was to determine which factors would be useful in predicting whether a user would become an adopted user (a user who logged in at least three times in an at least one seven day span).

The process of cleaning and restructuring the data was perhaps the most time consuming part of this challenge. For the 'users' dataset (the dataset that contained user names and emails) , we had to reformat the last_session_ creation_time to datetime format, delete the user names, and shorten the email names so that only the email brand remained. The reasoning behind this last decision was that we wanted to categorize the email data because perhaps users of one brand may be more likely to become adopted users.

We also turned the invited_by_user_id column into a binary variable in which we only state whether a user was invited by another user or not. Initially, almost half the values in this column were missing and no invitees had dramatically higher numbers than the rest.

For the user engagement dataset, we got rid of the visited column because it was redundant; the user_id column gave us the same basic info. Using resampling by looking for at least three user occurrences over any seven day period, we were able to determine which users were adopted users and create an adopted users column..

We then outer merged both datasets into through user_id. Several users didn't have last login times and the very same users didn't have any info regarding whether they were adopted users, so we dropped every such instance. Not knowing next to anything about the company's data collection process, we didn't know whether the records were simply incomplete or whether these users definitely were not adopted users. We erred on the side of caution by dropping them., even if it meant dropping about a fourth of the data. We then did a bit of feature engineering by subtracting the creation and last creation columns from one another and creating a new column - account_time. This would give us the duration of a user's account. Finally, we categorized the columns that had string based categorical data using get_dummies.

After the cleaning was complete, we introduced several machine learning models including logistic regression, random forest, SVC, and KNN. Using these models, we attempted to predict whether a user was an adopted user. Several of the models gave around 96% test accuracy. We optimized the hyperparameters for one of them - random forest- and were able to increase the test accuracy by a few hundredths of a percent. Using feature importance, we were able to determine that our newly created variable account_time was by far the most important factor, accounting for over 98% of the accuracy value.

One possible explanation is that the larger the difference between the last login and the original account creation, the more likely it is that the user is a repeat customer of the service. If the difference in times is relatively short, then it's more likely that they used the service once or twice and moved on to possibly another service.

If we wish to go more in depth, further research may include turning 'creation_time' into categorical data (morning, afternoon, and night, for example) to see if the time of day that an account was created would be a factor in predicting adoption.

One other approach that we didn't take was inferential analysis. We could analyze whether the difference in time between account creation and login between adopted users and nonadopted users is statistically significant. If it is, that would further support the idea that adopted users have had their accounts longer than nonadopted users.