Relax Challenge

Objective: To analyze a dataset of users who joined a product collected over the past two years and determine which factors are strong identifiers of a user being “adopted” which is defined as logging in to the product at least three times in any 7 day period.

Business Motivation: Being able to identify what factors lead to a user being adopted translates directly to where to target resources in order to increase ‘adoption rate’. For example, resources could be allocated more towards promoting a certain aspect of the product that is important to adoption rate or improve the aspects of the product that are lacking depending on the organization’s marketing strategy.

Data Analysis and Cleaning:

- Missing Data: We observe over 3000 samples with missing values for the ‘last_session_creation_time’ column and determine that there is no appropriate method of imputing so we omit this data.
- Adopted Users: The adopted users are not explicitly identified so we determine this by performing rolling window calculations over the takehome_user_engagement table.
- Categorical Variables: We have a categorical feature in ‘creation_source’ which we need to pre-process and convert encode as features that can be used in our classification model.

Machine Learning Classification: We could choose to use any classification algorithm as we are working with a trivial dataset so we arbitrarily choose to use a random forest classifier here. We also determine which features to use and hence we remove the user identifiers such as user ID and email address as these don’t provide any insight in predicting whether the user will be adopted or not.

We split into train-test split of 80/20 respectively and train our model and test the performance of our model with the following results:

Precision: 0.679
Recall: 0.587
F1 Score: 0.629

We then look at the importance of the features that we used for training and see that last_session_creation_time was our feature with the most predictive power.

Future Scope: Since we observe that last_session_creation_time is the feature with the most predictive power, we can extend our study by analyzing this aspect more systematically. For example, we could identify if there is any time in the day (morning, afternoon, evening, etc.) that users log in that lead to an increased adoption rate. Perhaps the month or season in which the account is created also correlates with adoption rate and the marketing team may use this information to advertise more during this time of year for example.

This project shows the importance of analyzing data and using insight from this analysis to develop your product.