**Relax Challenge**

I discovered that the last\_session\_creation\_time feature and the invited\_by\_user\_id feature had some missing data. The invited user column missing values were not filled, but rather replaced with a new feature based on if they were invited or not. The last session time was set equal to the initial creation time for those missing values. No other issues were noted with the data.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 1: Missing data by column

For feature engineering, the company name was stripped from the email address and put into 1 of 7 categories (top 6 companies or other), and to reduce the number of dates/ timestamps a time between the last session and the creation date was created while also truncating the creation date to only a month and year. The email address, name, last session date, and organization id features were all deleted because they were unique identifiers. The object id was also eliminated after it was used to join the predictor feature table to the adopted user data table.

Text

Description automatically generated

Figure 2: Top 6 email address companies

To prepare for the model, the categorical features were one-hot encoded, and then the data was split into training and test data such that 25% of the data was reserved for testing. A random forest model was chosen with default hyperparameters since it is typically a good out-of-the-bag predictor model. The model scores were an AUC of 1.0 and model accuracy of 0.97. The ranked list of feature importance was essentially only a single feature, with the time between the last session and the profile creation date having a weight of 0.886. I found this highly suspicious and realized that although the feature is not technically confounded with the target variable, the longer someone has had their account increases the likelihood that there was a week that they used the software three times. This information also is not particularly helpful for the company. A model was then ran without that feature, but the model accuracy was worse than if we predicted that everyone was not an adopted user.

A screenshot of a computer screen

Description automatically generated with medium confidence

Figure 3: Model performance with time between last session and creation

A screenshot of a computer screen

Description automatically generated with medium confidence

Figure 4: Model performance without time between last session and creation