RELAX CHALLENGE BRIEF REPORT

Data Cleaning:
I started out by categorizing the data into the correct data types.

- Missing values: I created new features that marked which values were missing, then dealt with the missing values appropriately
    - I decided to split the *invited_by_user_id* category into invited by superusers (defined as users who invited 5 or more people) or not.
    - To deal with the missing values of the *last_session_creation_time*, I filled them in with the *creation_time*. As a caveat, this could be a mistake if the users ended up creating their accounts within the period qualified for adoption, but wouldn't actually be considered as 'adopted' since they never logged into their accounts.
- Duplicates: I checked for duplicates and there were none.
- One-hot encoding: I one-hot encoded the categorical features.
- Feature engineering: On top of the superuser invitation feature, I added a few more features, breaking down the *creation_time* and *last_session_creation_time* into months, years, and days.
    - In hindsight, it might not have been a good idea to create *last_session_creation_time* features since those are directly factored into the definitions of what qualifies a user as adopted.
- Target variable: I then found the target variable by counting the maximum of a user's logins within a 7-day period. At the end, there weren't many users who met the target requirement: only 1597 out of 10403.

EDA:

- I plotted a heatmap of the correlations and there weren't many values that were correlated with the adoption target variable. Only *last_session_creation_time_isna* and *last_session_creation_year* were somewhat correlated, and I explained my issues with those above.
- There may have been some correlated variables not easily seen since I decided to drop the first variables when one-hot encoding, as to not create dependent variables in modeling.

Modeling:

- One quick random forest was created which performed better than the benchmark
- The model looked mainly at the last session creation dates and creation dates, so I'm not confident in the modeling section since the logic behind it may be faulty.