**Relax Inc. Data Challenge**

**Data Exploration / Feature Engineering**

To check for errors, I looked at each dataset for duplicates, inconsistent labeling, and numerical outliers. Most of the work came in feature engineering.

From the engagement data, I resampled the data into daily vists per user and using the rolling() function to count the number of days logged-in for each 7-day window. Grouping this data down to the max 7-day logins allowed me to flag each user as ‘engaged’ or not. I also used thi data to count the number of days between the user’s creation date and second login – hoping this could be a good indicator of future adoption.

For the user data, I broke out a new table of just the referral user ids to flag each user as having referred another user or not. I also calculated an account age. And finally merging all of these features into a final dataframe. Exploring our numerical features, the second login differential had some outliers, so I dropped any records >= 50 days. My logic being users that wait 50 days for another login aren’t representative of our larger user base.

A picture containing text, diagram, screenshot, line

Description automatically generated

Figure 1: Numerical Features Histograms

**Modeling**

I tested two models: Logistic Regression and XGBoost, using a pre-processing pipeline and GridSearchCV for hyperparameter tuning. For scoring I prioritized accuracy because I want our engagement predictions to not have a high rate of false positives. My Logistic Regression model had high accuracy (93%) but only 69% precision due to the imbalance classes.

My XGBoost returned

**Future Improvements**

I prioritized a simpler model for this exercise, but I think the primary improvement that could be made is addressing the class imbalance through SMOTE or another technique.