Matt Tilley

Relax Data Challenge

Cleaning Process:

The first step in the cleaning process was to take the engagement and users dataframe and merge them into one dataframe based on the user\_id column from the engagement table and object\_id from the takehome\_users dataframe. Once these two dataframes were merged, I filled in the missing values with 0 to account for there not being any data in the referral column.

Next, I created two new columns. The first was if a customer was referred by a friend or not. If they were they received a 1 in the invited column. If they were not then they received a 0 in this column. The second column I created was the Adopted User column. This column had a 1 if the user had been

Correlation

|  |  |  |  |
| --- | --- | --- | --- |
| Visited | Correlation | Adopted User | Correlation |
| Adopted User | 0.843 | Visited | 0.843 |
| Last\_session | 0.348 | Last\_session | 0.346 |
| Org\_id | 0.045 | Org\_id | 0.034 |
| Signup Google | 0.036 | Google | 0.021 |

It doesn’t look like there are any significant correlations other than adopted users visit the most so that is a good metric to track.

Information Value

I also ran a formula to determine the information value of each column. Unfortunately, this led to there not being any valuable prediction variables

Random Forest Important Features:

I also ran a random forest analysis with Adopted User being the target variable. By creating the model, I was able to generate the following important features:

|  |  |
| --- | --- |
| Feature | Importance |
| Visited | 0.986 |
| Guest Invite | 0.0002 |
| Org Invite | 0.0002 |
| Invited\_by user\_id | 0.0002 |
| Google | 0.0002 |

It doesn’t look like this revealed any more important features