**Missing Values:**

here are two features with missing values and I decided to remove them from the training set.

"last\_session\_creation\_time" is unix timestamp of last login. Originally was restored as float and after converting its data type to datetime, it represented bogus values.

As for "invited\_by\_user\_id", there can't be any values for two categories of "SIGNUP" and "SIGNUP\_GOOGLE\_AUTH". That's around 29% of our data.

**Visualization:**

The only two features that are correlated are “enabled\_for\_marketing\_drip” and “opted\_in\_to\_mailing\_list”.

The largest number of members are those who were invited by an organization. Also, members who were invited by an organization have higher tendency of subscribing to marketing emails.

**Predictive Models:**

Null Accuracy, the accuracy that could be achieved by always predicting the most frequent class for our dataset is 0.875 and none of the trained classification models such as Random Forest, SVM, Naïve Bayes, Logistic Regression could out perform it. The only model that performed slightly better than Null Accuracy was tuned GradientBoostingClassifier ( ‘learning\_rate’ : 0.01 , ‘max\_depth’ : 1 , ‘n\_estimator’ : 40 )with the accuracy of 0.877. I believe with the current data set we can’t get any model that outperform Null Accuracy.