**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:**

There was a huge imbalance between two classes, that means in order to evaluate the model’s performances we need to check confusion matrix, recall or precision, depending on the problem. In this case False Negative is more sensitive that False Positive, so we have to focus on lowering number False Negatives or increasing Recall value.

Initially all the models had a tendency of mislabeling most if not all the minority nodes as a majority class. After tuning hyper-parameters the best result with the lowest number of False Negatives was obtained from SVM model with C= 0.01 and gamma = 0.001.

It was no brainer in order to improve our result we needed to practice some common approaches for dealing with imbalanced datasets. Our preferences were over sampling of minority class and under sampling the majority class.

Comparing the result of Over sampling and under sampling on our best classifier(SVM), showed us under sampling with a high margin is performing much better than over sampling.

Here it is finally we got our best optimized model from combining under sampling with SVM.

Confucion Matrix:

[ 470 2997]

[ 55 438]]

Recall Score:

0.8884381338742393