There are 1602 adopted users and 10398 users that are not adopted in the entire dataset.

1) Correlation

Using Chi-squared test to check the pair-wise correlations between each predicting factor and the target variable ("adopted user"), the testing results are shown in the table below.

Factor	p-value	Significant at 5%-level?
creation_source	4.3058217090024059e-19	**Yes**
last_session_creation_time	2.7058998690135142e-103	**Yes**
opted_in_to_mailing_list	0.3919861918708256	No
enabled_for_marketing_drip	0.63691642256059933	No
org_id	4.2975282454675549e-05	**Yes**
invited_by_user_id	0.0022670517274883751	**Yes**

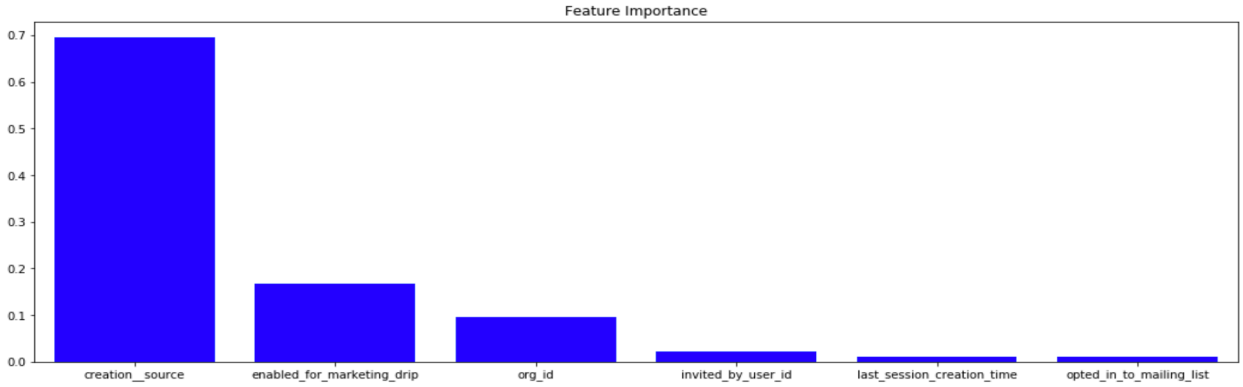
Therefore, at the 5% significance level, factors creation_source, last_session_creation_time, org_id, and invited_by_user_id are correlated with "adopted user".

2) Feature Importance

Splitting the data with stratification into 75% training and 25% test, we have

	Adopted User	Non-adopted User
Training	7798	1202
Test	2600	400

Training a random forest with 50 trees, the importance of the features is shown below



Hence, the useful factors in predicting user adoption are creation_source, enabled_for_marketing_drip. and org_id, where creation_source is the most important. Using the trained random forest on the test data, the recall, precision, and accuracy scores are 0.59, 0.76, and 0.92, respectively.

To further investigate the importance of these factors, the following should be considered:

- Tune the parameters of random forest, e.g. number of trees in the forest and number of features to try at each split
- Since the data is imbalanced (only 13.3% adopted users), resampling approaches (undersampling and oversampling) may potential boost the model performance and produce a more precise list of useful factors