Based on the information from about 12,000 users, a binary classifier was trained to predict the number weather a certain user is active or not. An active user was defined as the one who signed at least 3 times in a week’s period on separate days. Various features were extracted from the given data set and some further derived based on the data patterns. Among all the features, the number of visits, the time interval between the sign-up (account creation) and last session, and the organization a user belonged to that the most impact on user being active or not. Users two joined the site online via webpage or google authentication has more chances on being active. Figure 1 shows the results of the relative feature importance for predicting the

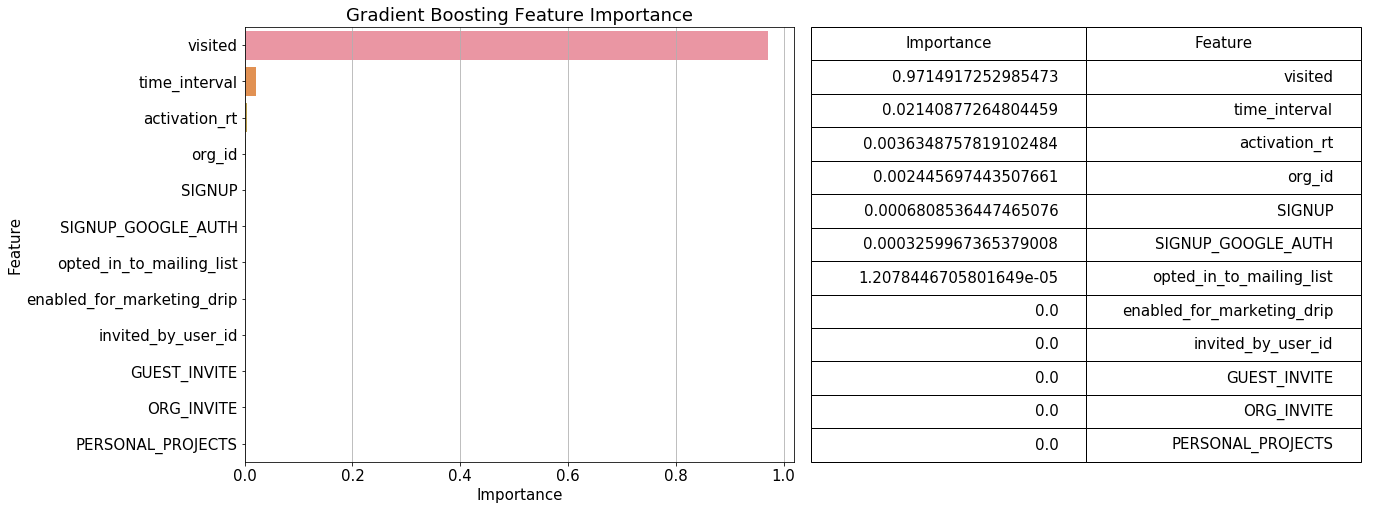


Figure 1. Feature importance

The dataset was trimmed down in order to have a relatively balanced class distribution (removing null values). This also helped in including relevant features (number of site visits and time interval) without oversimplification (filling in null). Figure 2 shows the confusion matrix. Despite the high class imbalance, the classifier was able to correctly label 448 (out of 487) as active, with F1-score of 94%.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted Labels | |
| 0 (Non-Active) | 1 (Active) |
| True Labels | 0 (Non-Active) | 2425 | 18 |
| 1 (Active) | 39 | 448 |

Figure 2. Confusion Matrix